Latent Variable Graphical Model Selection via Convex Optimization

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky *

Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

August 6, 2010

Abstract

Suppose we have samples of a *subset* of a collection of random variables. No additional information is provided about the number of latent variables, nor of the relationship between the latent and observed variables. Is it possible to discover the number of hidden components, and to learn a statistical model over the entire collection of variables? We address this question in the setting in which the latent and observed variables are jointly Gaussian, with the conditional statistics of the observed variables conditioned on the latent variables being specified by a graphical model. As a first step we give natural conditions under which such latent-variable Gaussian graphical models are identifiable given marginal statistics of only the observed variables. Essentially these conditions require that the conditional graphical model among the observed variables is sparse, while the effect of the latent variables is "spread out" over most of the observed variables. Next we propose a tractable convex program based on regularized maximum-likelihood for model selection in this latent-variable setting; the regularizer uses both the ℓ_1 norm and the nuclear norm. Our modeling framework can be viewed as a combination of dimensionality reduction (to identify latent variables) and graphical modeling (to capture remaining statistical structure not attributable to the latent variables), and it consistently estimates both the number of hidden components and the conditional graphical model structure among the observed variables. These results are applicable in the high-dimensional setting in which the number of latent/observed variables grows with the number of samples of the observed variables. The geometric properties of the algebraic varieties of sparse matrices and of low-rank matrices play an important role in our analysis.

Keywords: Gaussian graphical models; covariance selection; latent variables; regularization; sparsity; low-rank; algebraic statistics; high-dimensional asymptotics

1 Introduction

Statistical model selection in the high-dimensional regime arises in a number of applications. In many data analysis problems in geophysics, radiology, genetics, climate studies, and image processing, the number of samples available is comparable to or even smaller than the number of

*Email: {venkatc,parrilo,willsky}@mit.edu. This work was supported in part by AFOSR grant FA9550-08-1-0180, in part under a MURI through AFOSR grant FA9550-06-1-0324, in part under a MURI through AFOSR grant FA9550-06-1-0303, and in part by NSF FRG 0757207.

variables. However, it is well-known that empirical statistics such as sample covariance matrices are not well-behaved when both the number of samples and the number of variables are large and comparable to each other (see [26]). Model selection in such a setting is therefore both challenging and of great interest. In order for model selection to be well-posed given limited information, a key assumption that is often made is that the underlying model to be estimated only has *a few degrees of freedom*. Common assumptions are that the data are generated according to a graphical model, or a stationary time-series model, or a simple factor model with a few latent variables. Sometimes geometric assumptions are also made in which the data are viewed as samples drawn according to a distribution supported on a low-dimensional manifold.

A model selection problem that has received considerable attention recently is the estimation of covariance matrices in the high-dimensional setting. As the sample covariance matrix is poorly behaved in such a regime [20, 26], some form of *regularization* of the sample covariance is adopted based on assumptions about the true underlying covariance matrix. For example approaches based on banding the sample covariance matrix [3] have been proposed for problems in which the variables have a natural ordering (e.g., times series), while "permutation-invariant" methods that use thresholding are useful when there is no natural variable ordering [4, 15]. These approaches provide consistency guarantees under various sparsity assumptions on the true covariance matrix. Other techniques that have been studied include methods based on shrinkage [24, 39] and factor analysis [16]. A number of papers have studied covariance estimation in the context of *Gaussian graphical model selection*. In a Gaussian graphical model the *inverse* of the covariance matrix, also called the concentration matrix, is assumed to be sparse, and the sparsity pattern reveals the conditional independence relations satisfied by the variables. The model selection method usually studied in such a setting is ℓ_1-regularized maximum-likelihood, with the ℓ_1 penalty applied to the entries of the inverse covariance matrix to induce sparsity. The consistency properties of such an estimator have been studied [22, 29, 32], and under suitable conditions [22, 29] this estimator is also "sparsistent", i.e., the estimated concentration matrix has the same sparsity pattern as the true model from which the samples are generated. An alternative approach to ℓ_1-regularized maximum-likelihood is to estimate the sparsity pattern of the concentration matrix by performing regression separately on each variable [27]; while such a method consistently estimates the sparsity pattern, it does not directly provide estimates of the covariance or concentration matrix.

In many applications throughout science and engineering, a challenge is that one may not have access to observations of all the relevant phenomena, i.e., some of the relevant variables may be hidden or unobserved. Such a scenario arises in data analysis tasks in psychology, computational biology, and economics. In general latent variables pose a significant difficulty for model selection because one may not know the number of relevant latent variables, nor the relationship between these variables and the observed variables. Typical algorithmic methods that try to get around this difficulty usually fix the number of latent variables as well as the structural relationship between latent and observed variables (e.g., the graphical model structure between latent and observed variables), and use the EM algorithm to fit parameters [11]. This approach suffers from the problem that one optimizes non-convex functions, and thus one may get stuck in sub-optimal local minima. An alternative method that has been suggested is based on a greedy, local, combinatorial heuristic that assigns latent variables to groups of observed variables, based on some form of clustering of the observed variables [14]; however, this approach has no consistency guarantees.

In this paper we study the problem of latent-variable graphical model selection in the setting where all the variables, both observed and hidden, are jointly Gaussian. More concretely let the covariance matrix of a finite collection of jointly Gaussian random variables $X_O \cup X_H$ be denoted by $\Sigma_{(O\ H)}$, where X_O are the observed variables and X_H are the unobserved, hidden variables. The marginal statistics corresponding to the observed variables X_O are given by the marginal covariance

matrix Σ_O, which is simply a submatrix of the full covariance matrix $\Sigma_{(O\ H)}$. However suppose that we parameterize our model by the concentration matrix $K_{(O\ H)} = \Sigma^{-1}_{(O\ H)}$, which as discussed above reveals the connection to graphical models. In such a parametrization, the *marginal concentration matrix* Σ_O^{-1} corresponding to the observed variables X_O is given by the Schur complement [19] with respect to the block K_H:

$$\tilde{K}_O = \Sigma_O^{-1} = K_O - K_{O,H}K_H^{-1}K_{H,O}.$$

Thus if we only observe the variables X_O, we only have access to Σ_O (or $\tilde{K}_O$). The two terms that compose $\tilde{K}_O$ above have interesting properties. The matrix K_O specifies the concentration matrix of the *conditional statistics* of the observed variables given the latent variables. If these conditional statistics are given by a sparse graphical model then K_O is *sparse*. On the other hand the matrix $K_{O,H}K_H^{-1}K_{H,O}$ serves as a *summary* of the effect of marginalization over the hidden variables H. This matrix has small rank if the number of latent, unobserved variables H is small relative to the number of observed variables O (the rank is equal to $|H|$). Therefore the marginal concentration matrix $\tilde{K}_O$ of the observed variables X_O is generally *not sparse* due to the additional low-rank term $K_{O,H}K_H^{-1}K_{H,O}$. Hence standard graphical model selection techniques applied directly to the observed variables X_O are not useful.

A modeling paradigm that infers the effect of the latent variables X_H would be more suitable in order to provide a simple explanation of the underlying statistical structure. Hence we *decompose* $\tilde{K}_O$ into the sparse and low-rank components, which reveals the conditional graphical model structure in the observed variables as well as the *number* of and effect due to the unobserved latent variables. Such a method can be viewed as a blend of principal component analysis and graphical modeling. In standard graphical modeling one would directly approximate a concentration matrix by a sparse matrix in order to learn a sparse graphical model. On the other hand in principal component analysis the goal is to explain the statistical structure underlying a set of observations using a small number of latent variables (i.e., approximate a covariance matrix as a low-rank matrix). In our framework based on decomposing a concentration matrix, we learn a graphical model among the observed variables *conditioned* on a few (additional) latent variables. Notice that in our setting these latent variables are *not* principal components, as the conditional statistics (conditioned on these latent variables) are given by a graphical model. Therefore we refer to these latent variables informally as *hidden components*.

Our first contribution in Section 3 is to address the fundamental question of *identifiability* of such latent-variable graphical models given the marginal statistics of only the observed variables. The critical point is that we need to tease apart the correlations induced due to marginalization over the latent variables from the conditional graphical model structure among the observed variables. As the identifiability problem is one of *uniquely* decomposing the sum of a sparse matrix and a low-rank matrix into the individual components, we study the algebraic varieties of sparse matrices and low-rank matrices. An important theme in this paper is the connection between the tangent spaces to these algebraic varieties and the question of identifiability. Specifically let $\Omega(K_O)$ denote the tangent space at K_O to the algebraic variety of sparse matrices, and let $T(K_{O,H}K_H^{-1}K_{H,O})$ denote the tangent space at $K_{O,H}K_H^{-1}K_{H,O}$ to the algebraic variety of low-rank matrices. Then the *statistical* question of identifiability of K_O and $K_{O,H}K_H^{-1}K_{H,O}$ given $\tilde{K}_O$ is determined by the *geometric* notion of *transversality* of the tangent spaces $\Omega(K_O)$ and $T(K_{O,H}K_H^{-1}K_{H,O})$. The study of the transversality of these tangent spaces leads us to natural conditions for identifiability. In particular we show that latent-variable models in which (1) the sparse matrix K_O has a small number of nonzeros per row/column, and (2) the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$ has row/column spaces that are not closely aligned with the coordinate axes, are identifiable. These two conditions have natural statistical interpretations. The first condition ensures that there are no densely-

connected subgraphs in the conditional graphical model structure among the observed variables X_O given the hidden components, i.e., that these conditional statistics are indeed specified by a sparse graphical model. Such statistical relationships may otherwise be mistakenly attributed to the effect of marginalization over some latent variable. The second condition ensures that the effect of marginalization over the latent variables is "spread out" over many observed variables; thus, the effect of marginalization over a latent variable is not confused with the conditional graphical model structure among the observed variables. In fact the first condition is often assumed in some papers on standard graphical model selection without latent variables (see for example [29]). We note here that question of parameter identifiability was recently studied for models with discrete-valued latent variables (i.e., mixture models, hidden Markov models) [1]. However, this work is not applicable to our setting in which both the latent and observed variables are assumed to be jointly Gaussian.

As our next contribution we propose a *regularized maximum-likelihood decomposition* framework to approximate a given sample covariance matrix by a model in which the concentration matrix decomposes into a sparse matrix and a low-rank matrix. A number of papers over the last several years have suggested that heuristics based on using the ℓ_1 norm are very effective for recovering sparse models [6, 12, 13]. Indeed such heuristics have been effectively used, as described above, for model selection when the goal is to estimate sparse concentration matrices. In her thesis [17] Fazel suggested a convex heuristic based on the nuclear norm for rank-minimization problems in order to recover low-rank matrices. This method generalized the previously studied trace heuristic for recovering low-rank positive semidefinite matrices. Recently several conditions have been given under which these heuristics provably recover low-rank matrices in various settings [7, 30]. Motivated by the success of these heuristics, we propose the following penalized likelihood method given a sample covariance matrix Σ_O^n formed from n samples of the observed variables:

$$\begin{aligned}(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L} \quad & -\ell(S-L;\Sigma_O^n) \quad + \quad \lambda_n\,(\gamma\|S\|_1 + \mathrm{tr}(L)) \\ \text{s.t.} \quad & S-L \succ 0, \quad L \succeq 0.\end{aligned} \tag{1.1}$$

Here ℓ represents the Gaussian log-likelihood function and is given by $\ell(K;\Sigma) = \log\det(K) - \mathrm{tr}(K\Sigma)$ for $K \succ 0$, where tr is the trace of a matrix and det is the determinant. The matrix $\hat{S}_n$ provides an estimate of K_O, which represents the conditional concentration matrix of the observed variables; the matrix $\hat{L}_n$ provides an estimate of $K_{O,H}K_H^{-1}K_{H,O}$, which represents the effect of marginalization over the latent variables. Notice that the regularization function is a combination of the ℓ_1 norm applied to S and the nuclear norm applied to L (the nuclear norm reduces to the trace over the cone of symmetric, positive-semidefinite matrices), with γ providing a tradeoff between the two terms. This variational formulation is a *convex optimization* problem. In particular it is a regularized max-det problem and can be solved in polynomial time using standard off-the-shelf solvers [36].

Our main result in Section 4 is a proof of the consistency of the estimator (1.1) in the high-dimensional regime in which both the number of observed variables and the number of hidden components are allowed to grow with the number of samples (of the observed variables). We show that for a suitable choice of the regularization parameter λ_n, there exists a range of values of γ for which the estimates $(\hat{S}_n, \hat{L}_n)$ have the same sparsity (and sign) pattern and rank as $(K_O, K_{O,H}(K_H)^{-1}K_{H,O})$ with high probability (see Theorem 4.1). The key technical requirement is an identifiability condition for the two components of the marginal concentration matrix $\tilde{K}_O$ with respect to the Fisher information (see Section 3.4). We make connections between our condition and the irrepresentability conditions required for support/graphical-model recovery using ℓ_1 regularization [29, 40]. Our results provide numerous scaling regimes under which consistency holds in latent-variable graphical model selection. For example we show that under suitable identifiability

conditions consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables (see Section 4.3).

Related previous work The problem of decomposing the sum of a sparse matrix and a low-rank matrix, with no additional noise, into the individual components was initially studied in [9] by a superset of the authors of the present paper. Specifically this work proposed a convex program using a combination of the ℓ_1 norm and the nuclear norm to recover the sparse and low-rank components, and derived conditions under which the convex program exactly recovers these components. In subsequent work Candès et al. [8] also studied this noise-free sparse-plus-low-rank decomposition problem, and provided guarantees for exact recovery using the convex program proposed in [9]. The problem setup considered in the present paper is quite different and is more challenging because we are only given access to an inexact sample covariance matrix, and we are interested in recovering components that preserve both the sparsity pattern and the rank of the components in the true underlying model. In addition to proving such a consistency result for the estimator (1.1), we also provide a statistical interpretation of our identifiability conditions and describe natural classes of latent-variable Gaussian graphical models that satisfy these conditions. As such our paper is closer in spirit to the many recent papers on covariance selection, but with the important difference that some of the variables are not observed.

Outline Section 2 gives some background on graphical models as well as the algebraic varieties of sparse and low-rank matrices. It also provides a formal statement of the problem. Section 3 discusses conditions under which latent-variable models are identifiable, and Section 4 states the main results of this paper. We provide experimental demonstration of the effectiveness of our estimator on synthetic and real data in Section 5. Section 6 concludes the paper with a brief discussion. The appendices include additional details and proofs of all of our technical results.

2 Background and Problem Statement

We briefly discuss concepts from graphical modeling and give a formal statement of the latent-variable model selection problem. We also describe various properties of the algebraic varieties of sparse matrices and of low-rank matrices. The following matrix norms are employed throughout this paper:

- $\|M\|_2$: denotes the spectral norm, which is the largest singular value of M.
- $\|M\|_\infty$: denotes the largest entry in magnitude of M.
- $\|M\|_F$: denotes the Frobenius norm, which is the square-root of the sum of the squares of the entries of M.
- $\|M\|_*$: denotes the nuclear norm, which is the sum of the singular values of M. This reduces to the trace for positive-semidefinite matrices.
- $\|M\|_1$: denotes the sum of the absolute values of the entries of M.

A number of *matrix operator* norms are also used. For example, let $\mathcal{Z} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be a linear operator acting on matrices. Then the induced operator norm $\|\mathcal{Z}\|_{q\to q}$ is defined as:

$$\|\mathcal{Z}\|_{q\to q} \triangleq \max_{N\in\mathbb{R}^{p\times p},\ \|N\|_q\leq 1} \|\mathcal{Z}(N)\|_q. \tag{2.1}$$

Therefore, $\|\mathcal{Z}\|_{F\to F}$ denotes the spectral norm of the matrix operator $\mathcal{Z}$. The only vector norm used is the Euclidean norm, which is denoted by $\|\cdot\|$.

2.1 Gaussian graphical models with latent variables

A graphical model [23] is a statistical model defined with respect to a graph $(V, \mathcal{E})$ in which the nodes index a collection of random variables $\{X_v\}_{v\in V}$, and the edges represent the conditional independence relations (Markov structure) among the variables. The absence of an edge between nodes $i, j \in V$ implies that the variables X_i, X_j are independent conditioned on all the other variables. A *Gaussian graphical model* (also commonly referred to as a Gauss-Markov random field) is one in which all the variables are jointly Gaussian [33]. In such models the sparsity pattern of the inverse of the covariance matrix, or the *concentration* matrix, directly corresponds to the graphical model structure. Specifically, consider a Gaussian graphical model in which the covariance matrix is given by $\Sigma \succ 0$ and the concentration matrix is given by $K = \Sigma^{-1}$. Then an edge $\{i, j\} \in \mathcal{E}$ is present in the underlying graphical model if and only if $K_{i,j} \neq 0$.

Our focus in this paper is on Gaussian models in which some of the variables may not be observed. Suppose O represents the set of nodes corresponding to observed variables X_O, and H the set of nodes corresponding to unobserved, hidden variables X_H with $O \cup H = V$ and $O \cap H = \emptyset$. The joint covariance is denoted by $\Sigma_{(O\ H)}$, and joint concentration matrix by $K_{(O\ H)} = \Sigma^{-1}_{(O\ H)}$. The submatrix Σ_O represents the marginal covariance of the observed variables X_O, and the corresponding marginal concentration matrix is given by the Schur complement with respect to the block K_H:

$$\tilde{K}_O = \Sigma_O^{-1} = K_O - K_{O,H}K_H^{-1}K_{H,O}. \tag{2.2}$$

The submatrix K_O specifies the concentration matrix of the conditional statistics of the observed variables conditioned on the hidden components. If these conditional statistics are given by a sparse graphical model then K_O is sparse. On the other hand the marginal concentration matrix $\tilde{K}_O$ of the marginal distribution of X_O is *not* sparse in general due to the extra correlations induced from marginalization over the latent variables X_H, i.e., due to the presence of the additional term $K_{O,H}K_H^{-1}K_{H,O}$. Hence, standard graphical model selection techniques in which the goal is to approximate a sample covariance by a sparse graphical model are not well-suited for problems in which some of the variables are hidden. However, the matrix $K_{O,H}K_H^{-1}K_{H,O}$ is a low-rank matrix if the number of hidden variables is much smaller than the number of observed variables (i.e., $|H| \ll |O|$). Therefore, a more appropriate model selection method is to approximate the sample covariance by a model in which the concentration matrix decomposes into the sum of a sparse matrix and a low-rank matrix. The objective here is to learn a sparse graphical model among the observed variables *conditioned* on some latent variables, as such a model explicitly accounts for the extra correlations induced due to unobserved, hidden components.

2.2 Problem statement

In order to analyze latent-variable model selection methods, we need to define an appropriate notion of model selection consistency for latent-variable graphical models. Notice that given the two components K_O and $K_{O,H}K_H^{-1}K_{H,O}$ of the concentration matrix of the marginal distribution (2.2), there are *infinitely* many configurations of the latent variables (i.e., matrices $K_H \succ 0, K_{O,H} = K_{H,O}^T$) that give rise to the *same* low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$. Specifically for any non-singular matrix $B \in \mathbb{R}^{|H|\times|H|}$, one can apply the transformations $K_H \to BK_HB^T, K_{O,H} \to K_{O,H}B^T$ and still preserve the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$. In *all* of these models the marginal statistics of the observed variables X_O remain the same upon marginalization over the latent variables

X_H. The key *invariant* is the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$, which *summarizes* the effect of marginalization over the latent variables. These observations give rise to the following notion of consistency:

Definition 2.1. *A pair of (symmetric) matrices (S, L) with $S, L \in \mathbb{R}^{|O|\times|O|}$ is an* algebraically consistent *estimate of a latent-variable Gaussian graphical model given by the concentration matrix $K_{(O\ H)}$ if the following conditions hold:*

1. *The sign-pattern of S is the same as that of K_O:*

$$\text{sign}(S_{i,j}) = \text{sign}((K_O)_{i,j}), \quad \forall i, j.$$

   Here we assume that $\text{sign}(0) = 0$.

2. *The rank of L is the same as the rank of $K_{O,H}K_H^{-1}K_{H,O}$:*

$$\text{rank}(L) = \text{rank}(K_{O,H}K_H^{-1}K_{H,O}).$$

3. *The concentration matrix $S - L$ can be realized as the marginal concentration matrix of an appropriate latent-variable model:*

$$S - L \succ 0, \qquad L \succeq 0.$$

The first condition ensures that S provides the correct structural estimate of the conditional graphical model (given by K_O) of the observed variables conditioned on the hidden components. This property is the same as the "sparsistency" property studied in standard graphical model selection [22, 29]. The second condition ensures that the number of hidden components is correctly estimated. Finally, the third condition ensures that the pair of matrices (S, L) leads to a realizable latent-variable model. In particular this condition implies that there exists a valid latent-variable model on $|O \cup H|$ variables in which (a) the conditional graphical model structure among the observed variables is given by S, (b) the number of latent variables $|H|$ is equal to the rank of L, and (c) the extra correlations induced due to marginalization over the latent variables is equal to L. Any method for matrix factorization (see for example, [38]) can be used to factorize the low-rank matrix L, depending on the properties that one desires in the factors (e.g., sparsity).

We also study parametric consistency in the usual sense, i.e., we show that one can produce estimates (S, L) that converge in various norms to the matrices $(K_O, K_{O,H}K_H^{-1}K_{H,O})$. Notice that proving (S, L) is close to $(K_O, K_{O,H}K_H^{-1}K_{H,O})$ in some norm does not in general imply that the support/sign-pattern and rank of (S, L) are the same as those of $(K_O, K_{O,H}K_H^{-1}K_{H,O})$. Therefore parametric consistency is different from algebraic consistency, which requires that (S, L) have the same support/sign-pattern and rank as $(K_O, K_{O,H}K_H^{-1}K_{H,O})$.

Goal Let $K^*_{(O\ H)}$ denote the concentration matrix of a Gaussian model. Suppose that we have n samples $\{X_O^i\}_{i=1}^n$ of the observed variables X_O. We would like to produce estimates $(\hat{S}_n, \hat{L}_n)$ that, with high-probability, are both algebraically consistent and parametrically consistent (in some norm).

2.3 Likelihood function and Fisher information

Given n samples $\{X^i\}_{i=1}^n$ of a finite collection of jointly Gaussian zero-mean random variables with concentration matrix K^*, we define the sample covariance as follows:

$$\Sigma^n \triangleq \frac{1}{n}\sum_{i=1}^n X_i X_i^T. \tag{2.3}$$

It is then easily seen that the log-likelihood function is given by:

$$\ell(K;\Sigma^n) = \log\det(K) - \mathrm{tr}(K\Sigma^n), \tag{2.4}$$

where $\ell(K;\Sigma_n)$ is a function of K. Notice that this function is strictly concave for $K \succ 0$. Now consider the latent-variable modeling problem in which we wish to model a collection of random variables X_O (with sample covariance Σ_O^n) by adding some extra variables X_H. With respect to the parametrization (S, L) (with S representing the conditional statistics of X_O given X_H, and L summarizing the effect of marginalization over the additional variables X_H), the likelihood function is given by:

$$\bar{\ell}(S, L; \Sigma_O^n) = \ell(S \ - \ L; \Sigma_O^n).$$

The function $\bar{\ell}$ is *jointly concave* with respect to the parameters (S, L) whenever $S - L \succ 0$, and it is this function that we use in our variational formulation (1.1) to learn a latent-variable model.

In the analysis of a convex program involving the likelihood function, the Fisher information plays an important role as it is the negative of the Hessian of the likelihood function and thus controls the curvature. As the first term in the likelihood function is linear, we need only study higher-order derivatives of the log-determinant function in order to compute the Hessian. Letting $\mathcal{I}$ denote the Fisher information matrix, we have that [5]

$$\mathcal{I}(K^*) \triangleq -\nabla_K^2 \log\det(K)|_{K=K^*} = (K^*)^{-1} \otimes (K^*)^{-1},$$

for $K^* \succ 0$. If K^* is a $p \times p$ concentration matrix, then the Fisher information matrix $\mathcal{I}(K^*)$ has dimensions $p^2 \times p^2$. Next consider the latent-variable situation with the variables indexed by O being observed and the variables indexed by H being hidden. The concentration matrix $\tilde{K}_O^* = (\Sigma_O^*)^{-1}$ of the marginal distribution of the observed variables O is given by the Schur complement (2.2), and the corresponding Fisher information matrix is given by

$$\mathcal{I}(\tilde{K}_O^*) = (\tilde{K}_O^*)^{-1} \otimes (\tilde{K}_O^*)^{-1} = \Sigma_O^* \otimes \Sigma_O^*.$$

Notice that this is precisely the $|O|^2 \times |O|^2$ submatrix of the full Fisher information matrix $\mathcal{I}(K^*_{(O\ H)}) = \Sigma^*_{(O\ H)} \otimes \Sigma^*_{(O\ H)}$ with respect to all the parameters $K^*_{(O\ H)} = (\Sigma^*_{(O\ H)})^{-1}$ (corresponding to the situation in which *all* the variables $X_{O\cup H}$ are observed). The matrix $\mathcal{I}(K^*_{(O\ H)})$ has dimensions $|O\cup H|^2 \times |O\cup H|^2$, while $\mathcal{I}(\tilde{K}_O^*)$ is an $|O|^2 \times |O|^2$ matrix. To summarize, we have for all $i, j, k, l \in O$ that:

$$\mathcal{I}(\tilde{K}_O^*)_{(i,j),(k,l)} = [\Sigma^*_{(O\ H)} \otimes \Sigma^*_{(O\ H)}]_{(i,j),(k,l)} = \mathcal{I}(K^*_{(O\ H)})_{(i,j),(k,l)}.$$

In Section 3.4 we impose various conditions on the Fisher information matrix $\mathcal{I}(\tilde{K}_O^*)$ under which our regularized maximum-likelihood formulation provides consistent estimates with high probability.

2.4 Algebraic varieties of sparse and low-rank matrices

An algebraic variety is the solution set of a system of polynomial equations. The set of sparse matrices and the set of low-rank matrices can be naturally viewed as algebraic varieties. Here we describe these varieties, and discuss some of their properties. Of particular interest in this paper are geometric properties of these varieties such as the tangent space and local curvature at a (smooth) point.

Let $\mathcal{S}(k)$ denote the set of matrices with at most k nonzeros:

$$\mathcal{S}(k) \triangleq \{M \in \mathbb{R}^{p\times p} \mid |\text{support}(M)| \leq k\}. \tag{2.5}$$

The set $\mathcal{S}(k)$ is an algebraic variety, and can in fact be viewed as a union of $\binom{p^2}{k}$ subspaces in $\mathbb{R}^{p\times p}$. This variety has dimension k, and it is smooth everywhere except at those matrices that have support size strictly smaller than k. For any matrix $M \in \mathbb{R}^{p\times p}$, consider the variety $\mathcal{S}(|\text{support}(M)|)$; M is a smooth point of this variety, and the tangent space at M is given by

$$\Omega(M) = \{N \in \mathbb{R}^{p\times p} \mid \text{support}(N) \subseteq \text{support}(M)\}. \tag{2.6}$$

In words the tangent space $\Omega(M)$ at a smooth point M is given by the set of all matrices that have support contained within the support of M. We view $\Omega(M)$ as a subspace in $\mathbb{R}^{p\times p}$.

Next let $\mathcal{L}(r)$ denote the algebraic variety of matrices with rank at most r:

$$\mathcal{L}(r) \triangleq \{M \in \mathbb{R}^{p\times p} \mid \text{rank}(M) \leq r\}. \tag{2.7}$$

It is easily seen that $\mathcal{L}(r)$ is an algebraic variety because it can be defined through the vanishing of all $(r+1)\times(r+1)$ minors. This variety has dimension equal to $r(2p-r)$, and it is smooth everywhere except at those matrices that have rank strictly smaller than r. Consider a rank-r matrix M with SVD given by $M = UDV^T$, where $U, V \in \mathbb{R}^{p\times r}$ and $D \in \mathbb{R}^{r\times r}$. The matrix M is a smooth point of the variety $\mathcal{L}(\text{rank}(M))$, and the tangent space at M with respect to this variety is given by

$$T(M) = \{UY_1^T + Y_2V^T \mid Y_1, Y_2 \in \mathbb{R}^{p\times r}\}. \tag{2.8}$$

In words the tangent space $T(M)$ at a smooth point M is the span of all matrices that have either the same row-space as M or the same column-space as M. As with $\Omega(M)$ we view $T(M)$ as a subspace in $\mathbb{R}^{p\times p}$.

In Section 3 we explore the connection between geometric properties of these tangent spaces and the identifiability problem in latent-variable graphical models.

2.5 Curvature of rank variety

The sparse matrix variety $\mathcal{S}(k)$ has the property that it has *zero* curvature at any smooth point. Consequently the tangent space at a smooth point M is the *same* as the tangent space at any point in a neighborhood of M. This property is implicitly used in the analysis of ℓ_1 regularized methods for recovering sparse models. The situation is more complicated for the low-rank matrix variety, because the curvature at any smooth point is nonzero. Therefore we need to study how the tangent space changes from one point to a neighboring point by analyzing how this variety curves locally. Indeed the amount of curvature at a point is directly related to the "angle" between the tangent space at that point and the tangent space at a neighboring point. For any subspace T of matrices, let $\mathcal{P}_T$ denote the projection onto T. Given two subspaces T_1, T_2 of the same dimension, we measure the "twisting" between these subspaces by considering the following quantity.

$$\rho(T_1, T_2) \triangleq \|\mathcal{P}_{T_1} - \mathcal{P}_{T_2}\|_{2\to 2} = \max_{\|N\|_2 \leq 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_2. \tag{2.9}$$

In Appendix A we briefly review relevant results from matrix perturbation theory; the key tool used to derive these results is the resolvent of a matrix [21]. Based on these tools we prove the following two results in Appendix B, which bound the twisting between the tangent spaces at nearby points. The first result provides a bound on the quantity ρ between the tangent spaces at a point and at its neighbor.

Proposition 2.1. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\sigma}{8}$. Further, let $M+\Delta$ be a rank-r matrix. Then we have that*

$$\rho(T(M+\Delta), T(M)) \leq \frac{2}{\sigma}\ \|\Delta\|_2.$$

The next result bounds the error between a point and its neighbor in the normal direction.

Proposition 2.2. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\| \leq \frac{\sigma}{8}$. Further, let $M+\Delta$ be a rank-r matrix. Then we have that*

$$\|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 \leq \frac{\|\Delta\|_2^2}{\sigma}.$$

These results suggest that the closer the smallest singular value is to zero, the more curved the variety is locally. Therefore we control the twisting between tangent spaces at nearby points by bounding the smallest nonzero singular value away from zero.

3 Identifiability

In the absence of additional conditions, the latent-variable model selection problem is ill-posed. In this section we discuss a set of conditions on latent-variable models that ensure that these models are identifiable given marginal statistics for a subset of the variables.

3.1 Structure between latent and observed variables

Suppose that the low-rank matrix that summarizes the effect of the hidden components is itself sparse. This leads to identifiability issues in the sparse-plus-low-rank decomposition problem. Statistically the additional correlations induced due to marginalization over the latent variables could be mistaken for the conditional graphical model structure of the observed variables. In order to avoid such identifiability problems the effect of the latent variables must be "diffuse" across the observed variables. To address this point the following quantity was introduced in [9] for any matrix M, defined with respect to the tangent space $T(M)$:

$$\xi(T(M)) \triangleq \max_{N\in T(M),\ \|N\|_2\leq 1} \|N\|_\infty. \tag{3.1}$$

Thus $\xi(T(M))$ being small implies that elements of the tangent space $T(M)$ cannot have their support concentrated in a few locations; as a result M cannot be too sparse. This idea is formalized in [9] by relating $\xi(T(M))$ to a notion of "incoherence" of the row/column spaces, where the row/column spaces are said to be incoherent with respect to the standard basis if these spaces are not aligned closely with any of the coordinate axes. Letting $M = UDV^T$ be the singular value decomposition of M, the incoherence of the row/column spaces of M (initially proposed and studied by Candès and Recht [7]) is defined as:

$$\mathrm{inc}(M) \triangleq \max\{\max_i \|P_U(e_i)\|, \max_i \|P_V(e_i)\|\}. \tag{3.2}$$

Here P_V, P_U denote projections[1] onto the row/column spaces of M, and e_i is the i'th standard basis vector. Hence $\text{inc}(M)$ measures the projection of the most "closely aligned" coordinate axis with the row/column spaces. For any rank-r matrix M we have that

$$\sqrt{\frac{r}{p}} \leq \text{inc}(M) \leq 1, \tag{3.3}$$

where the lower bound is achieved (for example) if the row/column spaces span any r columns of a $p \times p$ orthonormal Hadamard matrix, while the upper bound is achieved if the row or column space contains a standard basis vector. Typically a matrix M with incoherent row/column spaces would have $\text{inc}(M) \ll 1$. The following result (proved in [9]) shows that the more incoherent the row/column spaces of M, the smaller is $\xi(M)$.

Proposition 3.1. *For any $M \in \mathbb{R}^{p \times p}$, we have that*

$$\text{inc}(M) \leq \xi(T(M)) \leq 2 \text{ inc}(M),$$

where $\xi(T(M))$ and $\text{inc}(M)$ are defined in (3.1) and (3.2).

Based on these concepts we roughly require that the low-rank matrix that summarizes the effect of the latent variables be *incoherent*, thereby ensuring that the extra correlations due to marginalization over the hidden components cannot be confused with the conditional graphical model structure of the observed variables. Notice that the quantity inc is not just a measure of the number of latent variables, but also of the overall effect of the correlations induced by marginalization over these variables.

Curvature and change in ξ: As noted previously an important technical point is that the algebraic variety of low-rank matrices is locally curved at any smooth point. Consequently the quantity ξ changes as we move along the low-rank matrix variety smoothly. The quantity $\rho(T_1, T_2)$ introduced in (2.9) also allows us to bound the variation in ξ as follows.

Lemma 3.2. *Let T_1, T_2 be two matrix subspaces of the same dimension with the property that $\rho(T_1, T_2) < 1$, where ρ is defined in (2.9). Then we have that*

$$\xi(T_2) \leq \frac{1}{1 - \rho(T_1, T_2)} \; [\xi(T_1) + \rho(T_1, T_2)].$$

This lemma is proved in Appendix B.

3.2 Structure among observed variables

An identifiability problem also arises if the conditional graphical model among the observed variables contains a densely connected subgraph. These statistical relationships might be mistaken as correlations induced by marginalization over latent variables. Therefore we need to ensure that the conditional graphical model among the observed variables is sparse. We impose the condition that this conditional graphical model must have small "degree", i.e., no observed variable is directly connected to too many other observed variables conditioned on the hidden components. Notice that bounding the degree is a more refined condition than simply bounding the total number of nonzeros

[1] We denote projections onto vector subspaces (defined by a matrix) by P, and projections onto matrix subspaces (defined by a general linear operator) by the calligraphic $\mathcal{P}$.

as the *sparsity pattern* also plays a role. In [9] the authors introduced the following quantity in order to provide an appropriate measure of the sparsity pattern of a matrix:

$$\mu(\Omega(M)) \triangleq \max_{N \in \Omega(M), \|M\|_\infty \leq 1} \|N\|_2. \tag{3.4}$$

The quantity $\mu(\Omega(M))$ being small for a matrix implies that the spectrum of any element of the tangent space $\Omega(M)$ is not too "concentrated", i.e., the singular values of the elements of the tangent space are not too large. In [9] it is shown that a sparse matrix M with "bounded degree" (a small number of nonzeros per row/column) has small $\mu(M)$.

Proposition 3.3. *Let $M \in \mathbb{R}^{p \times p}$ be any matrix with at most $\deg_{\max}(M)$ nonzero entries per row/column, and with at least $\deg_{\min}(M)$ nonzero entries per row/column. With $\mu(\Omega(M))$ as defined in (3.4), we have that*

$$\deg_{\min}(M) \leq \mu(\Omega(M)) \leq \deg_{\max}(M).$$

3.3 Transversality of tangent spaces

Suppose that we have the sum of two vectors, each from two known subspaces. It is possible to uniquely recover the individual vectors from the sum if and only if the subspaces have a transverse intersection, i.e., they only intersect at the origin. This simple observation leads to an appealing algebraic notion of identifiability. Consider the situation in which we have the sum of a sparse matrix and a low-rank matrix. In addition to this sum, suppose that we are also given the tangent spaces at these matrices with respect to the algebraic varieties of sparse and low-rank matrices respectively. Then a necessary and sufficient condition for *local* identifiability is that these tangent spaces have a transverse intersection. It turns out that these transversality conditions on the tangent spaces are also sufficient for the regularized maximum-likelihood convex program (1.1) to provide consistent estimates of the number of hidden components and the conditional graphical model structure of the observed variables conditioned on the latent variables (without any side information about the tangent spaces).

In order to quantify the level of transversality between the tangent spaces Ω and T we study the *minimum gain* with respect to some norm of the addition operator restricted to the cartesian product $\mathcal{Y} = \Omega \times T$. More concretely let $\mathcal{A} : \mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p} \to \mathbb{R}^{p \times p}$ represent the addition operator, i.e., the operator that adds two matrices. Then given any matrix norm $\|\cdot\|_q$ on $\mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p}$, the minimum gain of $\mathcal{A}$ restricted to $\mathcal{Y}$ is defined as follows:

$$\epsilon(\Omega, T, \|\cdot\|_q) \triangleq \min_{(S,L) \in \Omega \times T,\ \|(S,L)\|_q = 1} \|\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{A} \mathcal{P}_{\mathcal{Y}}(S, L)\|_q,$$

where $\mathcal{P}_{\mathcal{Y}}$ denotes the projection onto the space $\mathcal{Y}$, and $\mathcal{A}^\dagger$ denotes the adjoint of the addition operator (with respect to the standard Euclidean inner-product). The tangent spaces Ω and T have a *transverse* intersection if and only if $\epsilon(\Omega, T, \|\cdot\|_q) > 0$. The "level" of transversality is measured by the magnitude of $\epsilon(\Omega, T, \|\cdot\|_q)$. Note that if the norm $\|\cdot\|_q$ used is the Frobenius norm, then $\epsilon(\Omega, T, \|\cdot\|_F)$ is the square of the *minimum singular value* of the addition operator $\mathcal{A}$ restricted to $\Omega \times T$.

A natural norm with which to measure transversality is the dual norm of the regularization function in (1.1), as the subdifferential of the regularization function is specified in terms of its dual. The reasons for this will become clearer as we proceed through this paper. Recall that the regularization function used in the variational formulation (1.1) is given by:

$$f_\gamma(S, L) = \gamma \|S\|_1 + \|L\|_*,$$

where the nuclear norm $\|\cdot\|_*$ reduces to the trace function over the cone of positive-semidefinite matrices. This function is a norm for all $\gamma > 0$. The dual norm of f_γ is given by

$$g_\gamma(S, L) = \max\left\{\frac{\|S\|_\infty}{\gamma}, \|L\|_2\right\}.$$

The following simple lemma records a useful property of the g_γ norm that is used several times throughout this paper.

Lemma 3.4. *Let Ω and T be tangent spaces at any points with respect to the algebraic varieties of sparse and low-rank matrices. Then for any matrix M, we have that $\|\mathcal{P}_\Omega(M)\|_\infty \leq \|M\|_\infty$ and that $\|\mathcal{P}_T(M)\|_2 \leq 2\|M\|_2$. Further we also have that $\|\mathcal{P}_{\Omega^\perp}(M)\|_\infty \leq \|M\|_\infty$ and that $\|\mathcal{P}_{T^\perp}(M)\|_2 \leq \|M\|_2$. Thus for any matrices M, N and for $\mathcal{Y} = \Omega \times T$, one can check that $g_\gamma(\mathcal{P}_\mathcal{Y}(M, N)) \leq 2g_\gamma(M, N)$ and that $g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}(M, N)) \leq g_\gamma(M, N)$.*

Next we define the quantity $\chi(\Omega, T, \gamma)$ as follows in order to study the transversality of the spaces Ω and T with respect to the g_γ norm:

$$\chi(\Omega, T, \gamma) \triangleq \max\left\{\frac{\xi(T)}{\gamma}, 2\mu(\Omega)\gamma\right\} \tag{3.5}$$

Here μ and ξ are defined in (3.4) and (3.1). We then have the following result (proved in Appendix C):

Lemma 3.5. *Let $S \in \Omega, L \in T$ be matrices such that $\|S\|_\infty = \gamma$ and let $\|L\|_2 = 1$. Then we have that $g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S, L)) \in [1 - \chi(\Omega, T, \gamma), 1 + \chi(\Omega, T, \gamma)]$, where $\mathcal{Y} = \Omega \times T$ and $\chi(\Omega, T, \gamma)$ is defined in (3.5). In particular we have that $1 - \chi(\Omega, T, \gamma) \leq \epsilon(\Omega, T, g_\gamma)$.*

The quantity $\chi(\Omega, T, \gamma)$ being small implies that the addition operator is essentially isometric when restricted to $\mathcal{Y} = \Omega \times T$. Stated differently the magnitude of $\chi(\Omega, T, \gamma)$ is a measure of the level of transversality of the spaces Ω and T. If $\mu(\Omega)\xi(T) < \frac{1}{2}$ then $\gamma \in (\xi(T), \frac{1}{2\mu(\Omega)})$ ensures that $\chi(\Omega, T, \gamma) < 1$, which in turn implies that the tangent spaces Ω and T have a transverse intersection.

Observation: Thus we have that the smaller the quantities $\mu(\Omega)$ and $\xi(T)$, the more transverse the intersection of the spaces Ω and T.

3.4 Conditions on Fisher information

The main focus of Section 4 is to analyze the regularized maximum-likelihood convex program (1.1) by studying its optimality conditions. The log-likelihood function is well-approximated in a neighborhood by a quadratic form given by the Fisher information (which measures the curvature, as discussed in Section 2.3). Let $\mathcal{I}^* = \mathcal{I}(\tilde{K}^*_O)$ denote the Fisher information evaluated at the true marginal concentration matrix $\tilde{K}^*_O = K^*_O - K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, where $K^*_{(O\ H)}$ represents the concentration matrix of the full model (see equation (2.2)). The appropriate measure of transversality between the tangent spaces[2] $\Omega = \Omega(K^*_O)$ and $T = T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O})$ is then in a space in which the inner-product is given by $\mathcal{I}^*$. Specifically, we need to analyze the minimum gain of the operator $\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}$ restricted to the space $\mathcal{Y} = \Omega \times T$. Therefore we impose several conditions on the Fisher information $\mathcal{I}^*$. We define quantities that control the gains of $\mathcal{I}^*$ restricted to Ω and T separately; these ensure that elements of Ω and elements of T are individually identifiable under the map $\mathcal{I}^*$. In addition we define quantities that, in conjunction with bounds on $\mu(\Omega)$ and $\xi(T)$, allow us to control the gain of $\mathcal{I}^*$ restricted to the direct-sum $\Omega \oplus T$.

[2]We implicitly assume that these tangent spaces are subspaces of the space of *symmetric* matrices.

$\mathcal{I}^*$ **restricted to** Ω: The minimum gain of the operator $\mathcal{P}_\Omega \mathcal{I}^* \mathcal{P}_\Omega$ restricted to Ω is given by

$$\alpha_\Omega \triangleq \min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega \mathcal{I}^* \mathcal{P}_\Omega(M)\|_\infty.$$

The maximum effect of elements in Ω in the orthogonal direction $\Omega^\perp$ is given by

$$\delta_\Omega \triangleq \max_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} \mathcal{I}^* \mathcal{P}_\Omega(M)\|_\infty.$$

The operator $\mathcal{I}^*$ is injective on Ω if $\alpha_\Omega > 0$. The ratio $\frac{\delta_\Omega}{\alpha_\Omega} \leq 1 - \nu$ implies the irrepresentability condition imposed in [29], which gives a sufficient condition for consistent recovery of graphical model structure using ℓ_1-regularized maximum-likelihood. Notice that this condition is a generalization of the usual Lasso irrepresentability conditions [40], which are typically imposed on the covariance matrix. Finally we also consider the following quantity, which controls the behavior of $\mathcal{I}^*$ restricted to Ω in the spectral norm:

$$\beta_\Omega \triangleq \max_{M \in \Omega, \|M\|_2 = 1} \|\mathcal{I}^*(M)\|_2.$$

$\mathcal{I}^*$ **restricted to** T: Analogous to the case of Ω one could control the gains of the operators $\mathcal{P}_{T^\perp} \mathcal{I}^* \mathcal{P}_T$ and $\mathcal{P}_T \mathcal{I}^* \mathcal{P}_T$. However as discussed previously one complication is that the tangent spaces at nearby smooth points on the rank variety are in general different, and the amount of twisting between these spaces is governed by the local curvature. Therefore we control the gains of the operators $\mathcal{P}_{T'^\perp} \mathcal{I}^* \mathcal{P}_{T'}$ and $\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}$ for all tangent spaces T' that are "close to" the nominal T (at the true underlying low-rank matrix), measured by $\rho(T, T')$ (2.9) being small. The minimum gain of the operator $\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}$ restricted to T' (close to T) is given by

$$\alpha_T \triangleq \min_{\rho(T', T) \leq \frac{\xi(T)}{2}} \min_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'} \mathcal{I}^* \mathcal{P}_{T'}(M)\|_2.$$

Similarly the maximum effect of elements in T' in the orthogonal direction $T'^\perp$ (for T' close to T) is given by

$$\delta_T \triangleq \max_{\rho(T', T) \leq \frac{\xi(T)}{2}} \max_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'^\perp} \mathcal{I}^* \mathcal{P}_{T'}(M)\|_2.$$

Implicit in the definition of α_T and δ_T is the fact that the outer minimum and maximum are only taken over spaces T' that are tangent spaces to the rank-variety. The operator $\mathcal{I}^*$ is injective on all tangent spaces T' such that $\rho(T', T) \leq \frac{\xi(T)}{2}$ if $\alpha_T > 0$. An irrepresentability condition (analogous to those developed for the sparse case) for tangent spaces near T to the rank variety would be that $\frac{\delta_T}{\alpha_T} \leq 1 - \nu$. Finally we also control the behavior of $\mathcal{I}^*$ restricted to T' close to T in the ℓ_∞ norm:

$$\beta_T \triangleq \max_{\rho(T', T) \leq \frac{\xi(T)}{2}} \max_{M \in T', \|M\|_\infty = 1} \|\mathcal{I}^*(M)\|_\infty.$$

The two sets of quantities $(\alpha_\Omega, \delta_\Omega)$ and (α_T, δ_T) essentially control how $\mathcal{I}^*$ behaves when restricted to the spaces Ω and T *separately* (in the natural norms). The quantities β_Ω and β_T are useful in order to control the gains of the operator $\mathcal{I}^*$ restricted to the *direct sum* $\Omega \oplus T$. Notice that although the magnitudes of elements in Ω are measured most naturally in the ℓ_∞ norm, the quantity β_Ω is specified with respect to the spectral norm. Similarly elements of the tangent spaces T' to the rank variety are most naturally measured in the spectral norm, but β_T provides control in the ℓ_∞ norm. These quantities, combined with $\mu(\Omega)$ and $\xi(T)$ (defined in (3.4) and (3.1)), provide

the "coupling" necessary to control the behavior of $\mathcal{I}^*$ restricted to elements in the direct sum $\Omega \oplus T$. In order to keep track of fewer quantities, we summarize the six quantities as follows:

$$\begin{aligned} \alpha &\triangleq \min(\alpha_\Omega, \alpha_T) \\ \delta &\triangleq \max(\delta_\Omega, \delta_T) \\ \beta &\triangleq \max(\beta_\Omega, \beta_T). \end{aligned}$$

Main assumption There exists a $\nu \in (0, \frac{1}{2}]$ such that:

$$\frac{\delta}{\alpha} \leq 1 - 2\nu.$$

This assumption is to be viewed as a generalization of the irrepresentability conditions imposed on the covariance matrix [40] or the Fisher information matrix [29] in order to provide consistency guarantees for sparse model selection using the ℓ_1 norm. With this assumption we have the following proposition, proved in Appendix C, about the gains of the operator $\mathcal{I}^*$ restricted to $\Omega \oplus T$. This proposition plays a fundamental role in the analysis of the performance of the regularized maximum-likelihood procedure (1.1).

Proposition 3.6. *Let Ω and T be the tangent spaces defined in this section, and let $\mathcal{I}^*$ be the Fisher information evaluated at the true marginal concentration matrix. Further let α, β, ν be as defined above. Suppose that*

$$\mu(\Omega)\xi(T) \leq \frac{1}{6}\left(\frac{\nu\alpha}{\beta(2-\nu)}\right)^2,$$

and that γ is in the following range:

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right].$$

Then we have the following two conclusions for $\mathcal{Y} = \Omega \times T'$ with $\rho(T', T) \leq \frac{\xi(T)}{2}$:

1. *The minimum gain of $\mathcal{I}^*$ restricted to $\Omega \oplus T'$ is bounded below:*

$$\min_{(S,L)\in\mathcal{Y},\ \|S\|_\infty=\gamma,\ \|L\|_2=1} g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \quad \geq \quad \frac{\alpha}{2}.$$

   Specifically this implies that for all $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \geq \frac{\alpha}{2} g_\gamma(S,L).$$

2. *The effect of elements in $\mathcal{Y} = \Omega \times T'$ on the orthogonal complement $\mathcal{Y}^\perp = \Omega^\perp \times T'^\perp$ is bounded above:*

$$\left\| \mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\left(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\right)^{-1} \right\|_{g_\gamma \to g_\gamma} \leq 1 - \nu.$$

   Specifically this implies that for all $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \leq (1-\nu) g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)).$$

The last quantity we consider is the spectral norm of the marginal covariance matrix $\Sigma_O^* = (\tilde{K}_O^*)^{-1}$:

$$\psi \triangleq \|\Sigma_O^*\|_2 = \|(\tilde{K}_O^*)^{-1}\|_2. \tag{3.6}$$

A bound on ψ is useful in the probabilistic component of our analysis, in order to derive convergence rates of the sample covariance matrix to the true covariance matrix. We also observe that

$$\|\mathcal{I}^*\|_{2\to 2} = \|(\tilde{K}_O^*)^{-1} \otimes (\tilde{K}_O^*)^{-1}\|_{2\to 2} = \psi^2.$$

4 Regularized Maximum-Likelihood Convex Program and Consistency

4.1 Setup

Let $K^*_{(O\ H)}$ denote the full concentration matrix of a collection of zero-mean jointly-Gaussian observed and latent variables, let $p = |O|$ denote the number of observed variables, and let $h = |H|$ denote the number of latent variables. We are given n samples $\{X^i_O\}_{i=1}^n$ of the observed variables X_O. We consider the high-dimensional setting in which (p, h, n) are all allowed to grow simultaneously. The quantities α, β, ν, ψ defined in the previous section are accounted for in our analysis, although we suppress the dependence on these quantities in the statement of our main result. We explicitly keep track of the quantities $\mu(\Omega(K^*_O))$ and $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$ as these control the complexity of the latent-variable model given by $K^*_{(O\ H)}$. In particular μ controls the sparsity of the conditional graphical model among the observed variables, while ξ controls the incoherence or "diffusivity" of the extra correlations induced due to marginalization over the hidden variables. Based on the tradeoff between these two quantities, we obtain a number of classes of latent-variable graphical models (and corresponding scalings of (p, h, n)) that can be consistently recovered using the regularized maximum-likelihood convex program (1.1) (see Section 4.3 for details). Specifically we show that consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables. We present our main result next demonstrating the consistency of the estimator (1.1), and then discuss classes of latent-variable graphical models and various scaling regimes in which our estimator is consistent.

4.2 Main results

Given n samples $\{X^i_O\}_{i=1}^n$ of the observed variables X_O, the sample covariance is defined as:

$$\Sigma^n_O = \frac{1}{n}\sum_{i=1}^n X^i_O (X^i_O)^T.$$

As discussed in Section 2.2 the goal is to produce an estimate given by a pair of matrices (S, L) of the latent-variable model represented by $K^*_{(O\ H)}$. We study the consistency properties of the following regularized maximum-likelihood convex program:

$$\begin{aligned}(\hat{S}_n, \hat{L}_n) = \arg\min_{S,L}\ & \mathrm{tr}[(S-L)\ \Sigma^n_O] - \log\det(S-L)\ +\ \lambda_n[\gamma\|S\|_1 + \mathrm{tr}(L)] \\ \text{s.t.}\quad & S - L \succ 0, \quad L \succeq 0.\end{aligned} \tag{4.1}$$

Here λ_n is a regularization parameter, and γ is a tradeoff parameter between the rank and sparsity terms. Notice from Proposition 3.6 that the choice of γ depends on the values of $\mu(\Omega(K^*_O))$ and $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$; essentially these quantities correspond to the degree of the conditional graphical model structure of the observed variables and the incoherence of the low-rank matrix summarizing the effect of the latent variables (see Section 3). While these quantities may not be known *a priori*, we discuss a method to choose γ numerically in our experimental results (see Section 5). The following theorem shows that the estimates $(\hat{S}_n, \hat{L}_n)$ provided by the convex program (4.1) are consistent for a suitable choice of λ_n. In addition to the appropriate identifiability conditions (as specified by Proposition 3.6), we also impose lower bounds on the minimum nonzero entry of the sparse conditional graphical model matrix K^*_O and on the minimum nonzero singular

value of the low-rank matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ summarizing the effect of the hidden variables. We suppress the dependence[3] on α, β, ν, ψ as we assume that these quantities remain bounded and do not scale with the other parameters. We emphasize the dependence on $\mu(\Omega(K^*_O))$ and $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$ because these control the complexity of the underlying latent-variable graphical model as discussed above.

Theorem 4.1. *Let $K^*_{(O\ H)}$ denote the concentration matrix of a Gaussian model. We have n samples $\{X^i_O\}_{i=1}^n$ of the p observed variables denoted by O. Let $\Omega = \Omega(K^*_O)$ and $T = T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O})$ denote the tangent spaces at K^*_O and at $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ with respect to the sparse and low-rank matrix varieties respectively.*

***Assumptions**: Suppose that the following conditions hold:*

1. *The quantities $\mu(\Omega)$ and $\xi(T)$ satisfy the assumption of Proposition 3.6 for identifiability, and γ is chosen in the range specified by Proposition 3.6.*

2. *The number of samples n available is such that*

$$n \gtrsim \frac{p}{\xi(T)^4}.$$

3. *The regularization parameter λ_n is chosen as*

$$\lambda_n \asymp \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

4. *The minimum nonzero singular value σ of $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ is bounded as*

$$\sigma \gtrsim \frac{1}{\xi(T)^3}\sqrt{\frac{p}{n}}.$$

5. *The minimum magnitude nonzero entry θ of K^*_O is bounded as*

$$\theta \gtrsim \frac{1}{\xi(T)\mu(\Omega)}\sqrt{\frac{p}{n}}.$$

***Conclusions**: Then with probability greater than $1 - 2\exp\{-p\}$ we have:*

1. *Algebraic consistency: The estimate $(\hat{S}_n, \hat{L}_n)$ given by the convex program (4.1) is algebraically consistent, i.e., the support and sign pattern of $\hat{S}_n$ is the same as that of K^*_O, and the rank of $\hat{L}_n$ is the same as that of $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$.*

2. *Parametric consistency: The estimate $(\hat{S}_n, \hat{L}_n)$ given by the convex program (4.1) is parametrically consistent:*

$$g_\gamma(\hat{S}_n - K^*_O, \hat{L}_n - K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}) \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

[3] We use the notation $a \gtrsim b$ if there exists a function $r(\alpha, \beta, \nu, \psi)$ such that $a \geq r(\alpha, \beta, \nu, \psi)b$. Similarly we use the notation $a \asymp b$ if there exists a function $r(\alpha, \beta, \nu, \psi)$ such that $a = r(\alpha, \beta, \nu, \psi)b$.

The proof of this theorem is given in Appendix D. The theorem essentially states that if the minimum nonzero singular value of the low-rank piece $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ and minimum nonzero entry of the sparse piece K^*_O are bounded away from zero, then the convex program (4.1) provides estimates that are both algebraically consistent and parametrically consistent (in the ℓ_∞ and spectral norms). In Section 4.4 we also show that these results easily lead to parametric consistency rates for the corresponding estimate $(\hat{S}_n - \hat{L}_n)^{-1}$ of the marginal covariance Σ^*_O of the observed variables.

Notice that the condition on the minimum singular value of $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ is more stringent than on the minimum nonzero entry of K^*_O. One role played by these conditions is to ensure that the estimates $(\hat{S}_n, \hat{L}_n)$ do not have smaller support size/rank than $(K^*_O, K^*_{O,H}(K^*_H)^{-1}K^*_{H,O})$. However the minimum singular value bound plays the additional role of bounding the curvature of the low-rank matrix variety around the point $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, which is the reason for this condition being more stringent. Notice also that the number of hidden variables h does not explicitly appear in the bounds in Theorem 4.1, which only depend on $p, \mu(\Omega(K^*_O)), \xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$. However the dependence on h is implicit in the dependence on $\xi(T(K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}))$, and we discuss this point in greater detail in the following section.

Finally we note that algebraic and parametric consistency hold under the assumptions of Theorem 4.1 for a *range* of values of γ:

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right].$$

In particular the assumptions on the sample complexity, the minimum nonzero singular value of $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, and the minimum magnitude nonzero entry of K^*_O are governed by the lower end of this range for γ. These assumptions can be weakened if we only require consistency for a smaller range of values of γ. The following corollary conveys this point with a specific example:

Corollary 4.2. *Consider the same setup and notation as in Theorem 4.1. Suppose that the quantities $\mu(\Omega)$ and $\xi(T)$ satisfy the assumption of Proposition 3.6 for identifiability. Suppose that we make the following assumptions:*

1. *Let γ be chosen to be equal to $\frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}$ (the upper end of the range specified in Proposition 3.6), i.e., $\gamma \asymp \frac{1}{\mu(\Omega)}$.*
2. *$n \gtrsim \mu(\Omega)^4\ p$.*
3. *$\lambda_n \asymp \mu(\Omega)\sqrt{\frac{p}{n}}$.*
4. *$\sigma \gtrsim \frac{\mu(\Omega)^2}{\xi(T)}\sqrt{\frac{p}{n}}$.*
5. *$\theta \gtrsim \sqrt{\frac{p}{n}}$.*

Then with probability greater than $1 - 2\exp\{-p\}$ we have estimates $(\hat{S}_n, \hat{L}_n)$ that are algebraically consistent, and parametrically consistent with the error bounded as

$$g_\gamma(\hat{S}_n - K^*_O, \hat{L}_n - K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}) \lesssim \mu(\Omega)\sqrt{\frac{p}{n}}.$$

The proof of this corollary[4] is analogous to that of Theorem 4.1. We emphasize that in practice it is often beneficial to have consistent estimates for a range of values of γ (as in Theorem 4.1). Specifically the stability of the sparsity pattern and rank of the estimates $(\hat{S}_n, \hat{L}_n)$ for a range of tradeoff parameters is useful in order to choose a suitable value of γ, as prior information about the quantities $\mu(\Omega(K_O^*))$ and $\xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$ is not typically available (see Section 5).

4.3 Scaling regimes

Next we consider classes of latent-variable models that satisfy the conditions of Theorem 4.1. Recall that n denotes the number of samples, p denotes the number of observed variables, and h denotes the number of latent variables. Recall the assumption that the quantities α, β, ν, ψ defined in Section 3.4 remain bounded, and do not scale with the other parameters such as (p, h, n) or $\xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$ or $\mu(\Omega(K_O^*))$. In particular we focus on the tradeoff between $\xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$ and $\mu(\Omega(K_O^*))$ (the quantities that control the complexity of a latent-variable graphical model), and the resulting scaling regimes for consistent estimation. Let $d = \deg(K_O^*)$ denote the degree of the conditional graphical model among the observed variables, and let $i = \text{inc}(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the incoherence of the correlations induced due to marginalization over the latent variables (we suppress the dependence on n). These quantities are defined in Section 3, and we have from Propositions 3.1 and 3.3 that

$$\mu(\Omega(K_O^*)) \leq d, \quad \xi(T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)) \leq 2i.$$

Since α, β, ν, ψ are assumed to be bounded, we also have from Proposition 3.6 that the product of μ and ξ must be bounded by a constant. Thus, we study latent-variable models in which

$$d\ i = \mathcal{O}(1).$$

As we describe next, there are non-trivial classes of latent-variable graphical models in which this condition holds.

Bounded degree and incoherence: The first class of latent-variable models that we consider are those in which the conditional graphical model among the observed variables (given by K_O^*) has constant degree d. Recall from equation (3.3) that the incoherence i of the effect of the latent variables (given by $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$) can be as small as $\sqrt{\frac{h}{p}}$. Consequently latent-variable models in which

$$d = \mathcal{O}(1), \quad h \sim p,$$

can be estimated consistently from $n \sim p$ samples as long as the low-rank matrix $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is almost maximally incoherent, i.e., $i \sim \sqrt{\frac{h}{p}}$ so the effect of marginalization over the latent variables is diffuse across almost all the observed variables. Thus consistent latent-variable model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables.

Polylogarithmic degree The next class of models that we study are those in which the degree d of the conditional graphical model of the observed variables grows poly-logarithmically with p.

[4] By making stronger assumptions on the Fisher information matrix $\mathcal{I}^*$, one can further remove the factor of $\xi(T)$ in the lower bound for σ. Specifically the lower bound $\sigma \gtrsim \mu(\Omega)^3\sqrt{\frac{p}{n}}$ suffices for consistent estimation if α_T, β_T bound the minimum/maximum gains of $\mathcal{I}^*$ for *all* matrices (rather than just those near T), and δ_T bounds the $\mathcal{I}^*$-inner-product for *all* pairs of orthogonal matrices (rather than just those near T and $T^\perp$).

Consequently, the incoherence i of the matrix $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ must decay as the inverse of poly-$\log(p)$. Using the fact that maximally incoherent low-rank matrices $K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ can have incoherence as small as $\sqrt{\frac{h}{p}}$, latent-variable models in which

$$d \sim \log(p)^q, \quad h \sim \frac{p}{\log(p)^{2q}},$$

can be consistently estimated as long as $n \sim p$ poly-$\log(p)$.

4.4 Rates for covariance matrix estimation

The main result Theorem 4.1 gives conditions under which we can consistently estimate the sparse and low-rank parts that compose the marginal concentration matrix $\tilde{K}^*_O$. Here we prove a corollary that gives rates for covariance matrix estimation, i.e., the quality of the estimate $(\hat{S}_n - \hat{L}_n)^{-1}$ with respect to the "true" marginal covariance matrix Σ^*_O.

Corollary 4.3. *Under the same conditions as in Theorem 4.1, we have with probability greater than* $1 - 2\exp\{-p\}$ *that*

$$g_\gamma(\mathcal{A}^\dagger[(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma^*_O]) \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

Specifically this implies that $\|(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma^*_O\|_2 \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$.

Proof: The proof of this lemma follows directly from duality. Based on the analysis in Appendix D (in particular using the optimality conditions of the modified convex program (D.8)), we have that

$$g_\gamma(\mathcal{A}^\dagger[(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma^n_O]) \leq \lambda_n.$$

We also have from the bound on the number of samples n that with probability greater than $1 - 2\exp\{-p\}$ (see Appendix D.7)

$$g_\gamma(\mathcal{A}^\dagger[\Sigma^*_O - \Sigma^n_O]) \lesssim \lambda_n$$

Based on the choice of λ_n in Theorem 4.1, we then have the desired bound. □

4.5 Proof strategy for Theorem 4.1

Standard results from convex analysis [31] state that $(\hat{S}_n, \hat{L}_n)$ is a minimum of the convex program (4.1) if the zero matrix belongs to the subdifferential of the objective function evaluated at $(\hat{S}_n, \hat{L}_n)$ (in addition to $(\hat{S}_n, \hat{L}_n)$ satisfying the constraints). The subdifferential of the ℓ_1 norm at a matrix M is given by

$$N \in \partial\|M\|_1 \quad \Leftrightarrow \quad \mathcal{P}_{\Omega(M)}(N) = \operatorname{sign}(M), \ \|\mathcal{P}_{\Omega(M)^\perp}(N)\|_\infty \leq 1.$$

For a symmetric positive semidefinite matrix M with SVD $M = UDU^T$, the subdifferential of the trace function restricted to the cone of positive semidefinite matrices (i.e., the nuclear norm over this set) is given by:

$$N \in \partial[\operatorname{tr}(M) + \mathbb{I}_{M \succeq 0}] \quad \Leftrightarrow \quad \mathcal{P}_{T(M)}(N) = UU^T, \ \mathcal{P}_{T(M)^\perp}(N) \preceq I,$$

where $\mathbb{I}_{M \succeq 0}$ denotes the characteristic function of the set of positive semidefinite matrices (i.e., the convex function that evaluates to 0 over this set and ∞ outside). The key point is that

elements of the subdifferential decompose with respect to the tangent spaces $\Omega(M)$ and $T(M)$. This decomposition property plays a critical role in our analysis. In particular it states that the optimality conditions consist of two parts, one part corresponding to the tangent spaces Ω and T and another corresponding to the normal spaces $\Omega^{\perp}$ and $T^{\perp}$.

Consider the optimization problem (4.1) with the additional (non-convex) constraints that the variable S belongs to the algebraic variety of sparse matrices and that the variables L belongs to the algebraic variety of low-rank matrices. While this new optimization problem is non-convex, it has a very interesting property. At a globally optimal solution (and indeed at any locally optimal solution) $(\tilde{S}, \tilde{L})$ such that $\tilde{S}$ and $\tilde{L}$ are smooth points of the algebraic varieties of sparse and low-rank matrices, the first-order optimality conditions state that the Lagrange multipliers corresponding to the additional variety constraints must lie in the *normal spaces* $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$. This fundamental observation, combined with the decomposition property of the subdifferentials of the ℓ_1 and nuclear norms, suggests the following high-level proof strategy:

1. Let $(\tilde{S}, \tilde{L})$ be the globally optimal solution of the optimization problem (4.1) with the additional constraints that (S, L) belong to the algebraic varieties of sparse/low-rank matrices; specifically constrain S to lie in $\mathcal{S}(|\text{support}(K_O^*)|)$ and constrain L to lie in $\mathcal{L}(\text{rank}(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*))$. Show first that $(\tilde{S}, \tilde{L})$ are smooth points of these varieties.

2. The first part of the subgradient optimality conditions of the original convex program (4.1) corresponding to components *on* the tangent spaces $\Omega(\tilde{S})$ and $T(\tilde{L})$ is satisfied. This conclusion can be reached because the additional Lagrange multipliers due to the variety constraints lie in the normal spaces $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$.

3. Finally show that the second part of the subgradient optimality conditions of (4.1) (without any variety constraints) corresponding to components in the normal spaces $\Omega(\tilde{S})^{\perp}$ and $T(\tilde{L})^{\perp}$ is also satisfied by $(\tilde{S}, \tilde{L})$.

Combining these steps together we show that $(\tilde{S}, \tilde{L})$ satisfy the optimality conditions of the *original convex program* (4.1). Consequently $(\tilde{S}, \tilde{L})$ is also the optimum of the convex program (4.1). As this estimate is also the solution to the problem with the variety constraints, the algebraic consistency of $(\tilde{S}, \tilde{L})$ can be directly concluded. We emphasize here that the variety-constrained optimization problem is used solely as an analysis tool in order to prove consistency of the estimates provided by the convex program (4.1). These steps describe our broad strategy, and we refer the reader to Appendix D for details. The key technical complication is that the tangent spaces at $\tilde{L}$ and $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ are in general different. We bound the twisting between these tangent spaces by using the fact that the minimum nonzero singular value of $K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is bounded away from zero (as assumed in Theorem 4.1 and using Proposition 2.1).

5 Simulation Results

In this section we give experimental demonstration of the consistency of our estimator (4.1) on synthetic examples, and its effectiveness in modeling real-world stock return data. Our choices of λ_n and γ are guided by Theorem 4.1. Specifically, we choose λ_n to be proportional to $\sqrt{\frac{p}{n}}$. For γ we observe that the support/sign-pattern and the rank of the solution $(\hat{S}_n, \hat{L}_n)$ are the same for a *range* of values of γ. Therefore one could solve the convex program (4.1) for several values of γ, and choose a solution in a suitable range in which the sign-pattern and rank of the solution are stable. In practical problems with real-world data these parameters may be chosen via cross-validation.

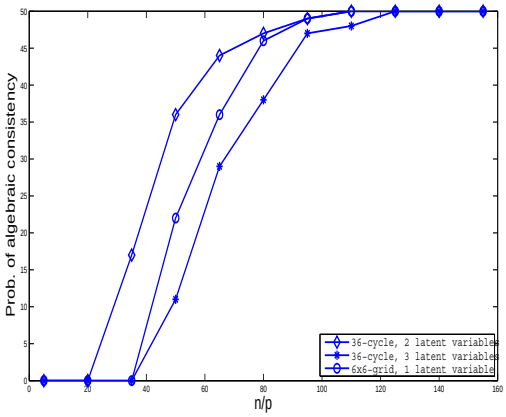

Figure 1: Synthetic data: Plot showing probability of consistent estimation of the number of latent variables, and the conditional graphical model structure of the observed variables. the three models studied are (a) 36-node conditional graphical model given by a cycle with $h = 2$ latent variables, (b) 36-node conditional graphical model given by a cycle with $h = 3$ latent variables, and (c) 36-node conditional graphical model given by a 6×6 grid with $h = 1$ latent variable. For each plotted point, the probability of consistent estimation is obtained over 50 random trials.

For small problem instances we solve the convex program (4.1) using a combination of YALMIP [25] and SDPT3 [34], which are standard off-the-shelf packages for solving convex programs. For larger problem instances we use the special purpose solver LogdetPPA [36] developed for log-determinant semidefinite programs.

5.1 Synthetic data

In the first set of experiments we consider a setting in which we have access to samples of the observed variables of a latent-variable graphical model. We consider several latent-variable Gaussian graphical models. The first model consists of $p = 36$ observed variables and $h = 2$ hidden variables. The conditional graphical model structure of the observed variables is a cycle with the edge partial correlation coefficients equal to 0.25; thus, this conditional model is specified by a sparse graphical model with degree 2. The second model is the same as the first one, but with $h = 3$ latent variables. The third model consists of $h = 1$ latent variable, and the conditional graphical model structure of the observed variables is given by a 6×6 nearest-neighbor grid (i.e., $p = 36$ and degree 4) with the partial correlation coefficients of the edges equal to 0.15. In all three of these models each latent variable is connected to a random subset of 80% of the observed variables (and the partial correlation coefficients corresponding to these edges are also random). Therefore the effect of the latent variables is "spread out" over most of the observed variables, i.e., the low-rank matrix summarizing the effect of the latent variables is incoherent.

For each model we generate n samples of the observed variables, and use the resulting sample covariance matrix Σ_O^n as input to our convex program (4.1). Figure 1 shows the probability of recovery of the support/sign-pattern of the conditional graphical model structure in the observed variables and the number of latent variables (i.e., probability of obtaining algebraically consistent estimates) as a function of n. This probability is evaluated over 50 experiments for each value of n.

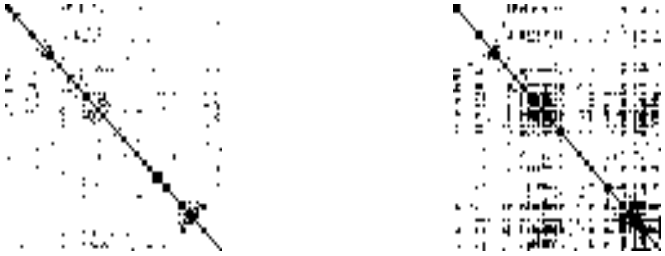

Figure 2: Stock returns: The figure on the left shows the sparsity pattern (black denotes an edge, and white denotes no edge) of the concentration matrix of the conditional graphical model (135 edges) of the stock returns, conditioned on 5 latent variables, in a latent-variable graphical model (total number of parameters equals 639). This model is learned using (4.1), and the KL divergence with respect to a Gaussian distribution specified by the sample covariance is 17.7. The figure on the right shows the concentration matrix of the graphical model (646 edges) of the stock returns, learned using standard sparse graphical model selection based on solving an ℓ_1-regularized maximum-likelihood program (total number of parameters equals 730). The KL divergence between this distribution and a Gaussian distribution specified by the sample covariance is 44.4.

In all of these cases standard graphical model selection applied directly to the observed variables is not useful as the marginal concentration matrix of the observed variables is not well-approximated by a sparse matrix. These experiments agree with our theoretical results that the convex program (4.1) is an algebraically consistent estimator of a latent-variable model given (sufficiently many) samples of only the observed variables.

5.2 Stock return data

In the next experiment we model the statistical structure of monthly stock returns of 84 companies in the S&P 100 index from 1990 to 2007; we disregard 16 companies that were listed after 1990. The number of samples n is equal to 216. We compute the sample covariance based on these returns and use this as input to (4.1).

The model learned using (4.1) for suitable values of λ_n, γ consists of $h = 5$ latent variables, and the conditional graphical model structure of the stock returns conditioned on these hidden components consists of 135 edges. Therefore the number of parameters in the model is $84 + 135 + (5 \times 84) = 639$. The resulting KL divergence between the distribution specified by this model and a Gaussian distribution specified by the sample covariance is 17.7. Figure 2 (left) shows the *conditional* graphical model structure. The strongest edges in this conditional graphical model, as measured by partial correlation, are between Baker Hughes - Schlumberger, A.T.&T. - Verizon, Merrill Lynch - Morgan Stanley, Halliburton - Baker Hughes, Intel - Texas Instruments, Apple - Dell, and Microsoft - Dell. It is of interest to note that in the Standard Industrial Classification[5] system for grouping these companies, several of these pairs are in different classes. As mentioned in Section 2.2 our method estimates a low-rank matrix that summarizes the effect of the latent variables; in order to factorize this low-rank matrix, for example into sparse factors, one could use methods such as those described in [38].

[5]See the United States Securities and Exchange Commission website at http://www.sec.gov/info/edgar/siccodes.htm

We compare these results to those obtained using a sparse graphical model learned using ℓ_1-regularized maximum-likelihood (see for example [29]), without introducing any latent variables. Figure 2 (right) shows this graphical model structure. The number of edges in this model is 646 (the total number of parameters is equal to $646 + 84 = 730$), and the resulting KL divergence between this distribution and a Gaussian distribution specified by the sample covariance is 44.4. Indeed to obtain a comparable KL divergence to that of the latent-variable model described above, one would require a graphical model with over 3000 edges.

These results suggest that a latent-variable graphical model is better suited than a standard sparse graphical model for modeling the statistical structure among stock returns. This is likely due to the presence of global, long-range correlations in stock return data that are better modeled via latent variables.

6 Discussion

We have studied the problem of modeling the statistical structure of a collection of random variables as a sparse graphical model conditioned on a few additional hidden components. As a first contribution we described conditions under which such latent-variable graphical models are identifiable given samples of only the observed variables. We also proposed a convex program based on regularized maximum-likelihood for latent-variable graphical model selection; the regularization function is a combination of the ℓ_1 norm and the nuclear norm. Given samples of the observed variables of a latent-variable Gaussian model we proved that this convex program provides consistent estimates of the number of hidden components as well as the conditional graphical model structure among the observed variables conditioned on the hidden components. Our analysis holds in the high-dimensional regime in which the number of observed/latent variables are allowed to grow with the number of samples of the observed variables. In particular we discuss certain scaling regimes in which consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables. These theoretical predictions are verified via a set of experiments on synthetic data. We also demonstrate the effectiveness of our approach in modeling real-world stock return data.

Several research questions arise that are worthy of further investigation. While the convex program (4.1) can be solved in polynomial time using off-the-shelf solvers, it is preferable to develop more efficient special-purpose solvers that can scale to massive datasets by taking advantage of the structure of the formulation (4.1). Finally it would be of interest to develop a similar convex optimization formulation with consistency guarantees for latent-variable models with non-Gaussian variables, e.g., for categorical data.

Acknowledgements

We would like to thank James Saunderson and Myung Jin Choi for helpful discussions, and Kim-Chuan Toh for kindly providing us specialized code to solve larger instances of our convex program.

A Matrix Perturbation Bounds

Given a low-rank matrix we consider what happens to the invariant subspaces when the matrix is perturbed by a small amount. We assume without loss of generality that the matrix under consideration is square and symmetric, and our methods can be extended to the general non-symmetric non-square case. We refer the interested reader to [2, 21] for more details, as the results

presented here are only a brief summary of what is relevant for this paper. In particular the arguments presented here are along the lines of those presented in [2]. The appendices in [2] also provide a more refined analysis of second-order perturbation errors.

The resolvent of a matrix M is given by $(M - \zeta I)^{-1}$ [21], and it is well-defined for all $\zeta \in \mathbb{C}$ that do not coincide with an eigenvalue of M. If M has no eigenvalue with magnitude equal to η, then we have by the Cauchy residue formula that the projector onto the invariant subspace of a matrix M corresponding to all singular values smaller than η is given by

$$P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} (M - \zeta I)^{-1} d\zeta, \tag{A.1}$$

where $\mathcal{C}_\eta$ denotes the positively-oriented circle of radius η centered at the origin. Similarly, we have that the weighted projection onto the invariant subspace corresponding to the smallest singular values is given by

$$P^w_{M,\eta} = M P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} \zeta \ (M - \zeta I)^{-1} d\zeta, \tag{A.2}$$

Suppose that M is a low-rank matrix with smallest nonzero singular value σ, and let Δ be a perturbation of M such that $\|\Delta\|_2 \leq \kappa < \frac{\sigma}{2}$. We have the following identity for any $|\zeta| = \kappa$, which will be used repeatedly:

$$[(M + \Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} = -[M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1}. \tag{A.3}$$

We then have that

$$\begin{aligned} P_{M+\Delta,\kappa} - P_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} [(M + \Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} [M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{A.4}$$

Similarly, we have the following for $P^w_{M,\kappa}$:

$$\begin{aligned} P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ \left\{ [(M + \Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ \left\{ [M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} d\zeta \\ &\quad - \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{A.5}$$

Given these expressions, we have the following two results.

Proposition A.1. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 \leq \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \ \|\Delta\|_2.$$

Proof: This result follows directly from the expression (A.4), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned}\|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 &\leq \frac{1}{2\pi}\ 2\pi\ \kappa\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma-\frac{3\kappa}{2}} \\ &= \frac{\kappa}{(\sigma-\kappa)(\sigma-\frac{3\kappa}{2})}\ \|\Delta\|_2.\end{aligned}$$

Here, we used the fact that $\|[M-\zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\kappa}$ and $\|[(M+\Delta)-\zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. $\square$

Next, we develop a similar bound for $P^w_{M,\kappa}$. Let $U(M)$ denote the invariant subspace of M corresponding to the nonzero singular values, and let $P_{U(M)}$ denote the projector onto this subspace.

Proposition A.2. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} - (I-P_{U(M)})\Delta(I-P_{U(M)})\|_2 \leq \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma-\frac{3\kappa}{2})}\ \|\Delta\|_2^2.$$

Proof: One can check that

$$\frac{1}{2\pi i}\oint_{\mathcal{C}_\kappa} \zeta\ [M-\zeta I]^{-1}\Delta[M-\zeta I]^{-1}d\zeta = (I-P_{U(M)})\Delta(I-P_{U(M)}).$$

Next we use the expression (A.5), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned}\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &- (I-P_{U(M)})\Delta(I-P_{U(M)})\|_2 \\ &\leq \frac{1}{2\pi}\ 2\pi\ \kappa\ \kappa\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma-\frac{3\kappa}{2}} \\ &= \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma-\frac{3\kappa}{2})}\ \|\Delta\|_2^2.\end{aligned}$$

As with the previous proof, we used the fact that $\|[M-\zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\kappa}$ and $\|[(M+\Delta)-\zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. $\square$

We will use these expressions to derive bounds on the "twisting" between the tangent spaces at M and at $M+\Delta$ with respect to the rank variety.

B Curvature of Rank Variety

For a symmetric rank-r matrix M, the projection onto the tangent space $T(M)$ (restricted to the variety of symmetric matrices with rank less than or equal to r) can be written in terms of the projection $P_{U(M)}$ onto the row space $U(M)$. For any matrix N

$$\mathcal{P}_{T(M)}(N) = P_{U(M)}N + NP_{U(M)} - P_{U(M)}NP_{U(M)}.$$

One can then check that the projection onto the normal space $T(M)^\perp$

$$\mathcal{P}_{T(M)^\perp}(N) = [I-\mathcal{P}_{T(M)}](N) = (I-P_{U(M)})\ N\ (I-P_{U(M)}).$$

Proof of Proposition 2.1: For any matrix N, we have that

$$\begin{aligned}[\mathcal{P}_{T(M+\Delta)} &- \mathcal{P}_{T(M)}](N) = \\ &[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I-P_{U(M)}] + [I-P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}].\end{aligned}$$

Further, we note that for $\kappa < \frac{\sigma}{2}$

$$\begin{aligned} P_{U(M+\Delta)} - P_{U(M)} &= [I - P_{U(M)}] - [I - P_{U(M+\Delta)}] \\ &= P_{M,\kappa} - P_{M+\Delta,\kappa}, \end{aligned}$$

where $P_{M,\kappa}$ is defined in the previous section. Thus, we have the following sequence of inequalities for $\kappa = \frac{\sigma}{4}$:

$$\begin{aligned} \rho(T(M+\Delta), T(M)) &= \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] \\ &\qquad + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}]\|_2 \\ &\quad + \max_{\|N\|_2 \leq 1} \|[I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq 2\ \|P_{M+\Delta,\frac{\sigma}{4}} - P_{M,\frac{\sigma}{4}}\|_2 \\ &\leq \frac{2}{\sigma}\ \|\Delta\|_2, \end{aligned}$$

where we obtain the last inequality from Proposition A.1. □

Proof of Proposition 2.2: Since both M and $M + \Delta$ are rank-r matrices, we have that $\mathcal{P}^w_{M+\Delta,\kappa} = \mathcal{P}^w_{M,\kappa} = 0$ for $\kappa = \frac{\sigma}{4}$. Consequently,

$$\begin{aligned} \|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 &= \|(I - P_{U(M)})\ \Delta\ (I - P_{U(M)})\|_2 \\ &\leq \frac{\|\Delta\|_2^2}{\sigma}, \end{aligned}$$

where we obtain the last inequality from Proposition A.2 with $\kappa = \frac{\sigma}{4}$. □

Proof of Lemma 3.2: Since $\rho(T_1, T_2) < 1$ one can check that the largest principal angle between T_1 and T_2 is strictly less than $\frac{\pi}{2}$. Consequently, the mapping $\mathcal{P}_{T_2} : T_1 \to T_2$ restricted to T_1 is bijective (as it is injective, and the spaces T_1, T_2 have the same dimension). Consider the maximum and minimum gain of the operator $\mathcal{P}_{T_2}$ restricted to T_1; for any $M \in T_1, \|M\|_2 = 1$:

$$\begin{aligned} \|\mathcal{P}_{T_2}(M)\|_2 &= \|M + [\mathcal{P}_{T_2} - \mathcal{P}_{T_1}](M)\|_2 \\ &\in [1 - \rho(T_1, T_2), 1 + \rho(T_1, T_2)]. \end{aligned}$$

Therefore, we can rewrite $\xi(T_2)$ as follows:

$$\begin{aligned} \xi(T_2) &= \max_{N \in T_2, \|N\|_2 \leq 1} \|N\|_\infty \\ &= \max_{N \in T_2, \|N\|_2 \leq 1} \|\mathcal{P}_{T_2}(N)\|_\infty \\ &\leq \max_{N \in T_1, \|N\|_2 \leq \frac{1}{1-\rho(T_1,T_2)}} \|\mathcal{P}_{T_2}(N)\|_\infty \\ &\leq \max_{N \in T_1, \|N\|_2 \leq \frac{1}{1-\rho(T_1,T_2)}} [\|N\|_\infty + \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_\infty] \\ &\leq \frac{1}{1 - \rho(T_1, T_2)} \left[\xi(T_1) + \max_{N \in T_1, \|N\|_2 \leq 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_\infty\right] \\ &\leq \frac{1}{1 - \rho(T_1, T_2)} \left[\xi(T_1) + \max_{\|N\|_2 \leq 1} \|[\mathcal{P}_{T_1} - \mathcal{P}_{T_2}](N)\|_2\right] \\ &\leq \frac{1}{1 - \rho(T_1, T_2)} \left[\xi(T_1) + \rho(T_1, T_2)\right]. \end{aligned}$$

This concludes the proof of the lemma. □

C Transversality and Identifiability

Proof of Lemma 3.5: We have that $\mathcal{A}^\dagger\mathcal{A}(S,L) = (S+L, S+L)$; therefore, $\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L) = (S+\mathcal{P}_\Omega(L), \mathcal{P}_T(S)+L)$. We need to bound $\|S+\mathcal{P}_\Omega(L)\|_\infty$ and $\|\mathcal{P}_T(S)+L\|_2$. First, we have

$$\begin{aligned}\|S+\mathcal{P}_\Omega(L)\|_\infty &\in [\|S\|_\infty - \|\mathcal{P}_\Omega(L)\|_\infty, \|S\|_\infty + \|\mathcal{P}_\Omega(L)\|_\infty] \\ &\subseteq [\|S\|_\infty - \|L\|_\infty, \|S\|_\infty + \|L\|_\infty] \\ &\subseteq [\gamma - \xi(T), \gamma + \xi(T)].\end{aligned}$$

Similarly, one can check that

$$\begin{aligned}\|\mathcal{P}_T(S)+L\|_2 &\in [-\|\mathcal{P}_T(S)\|_2 + \|L\|_2, \|\mathcal{P}_T(S)\|_2 + \|L\|_2] \\ &\subseteq [1-2\|S\|_2, 1+2\|S\|_2] \\ &\subseteq [1-2\gamma\mu(\Omega), 1+2\gamma\mu(\Omega)].\end{aligned}$$

Thus, we can conclude that

$$g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) \in [1-\chi(\Omega,T,\gamma), 1+\chi(\Omega,T,\gamma)].$$

where $\chi(\Omega,T,\gamma)$ is defined in (3.5). □

Proof of Proposition 3.6: Before proving the two parts of this proposition we make a simple observation about $\xi(T')$ using the condition that $\rho(T,T') \leq \frac{\xi(T)}{2}$ by applying Lemma 3.2:

$$\begin{aligned}\xi(T') &\leq \frac{\xi(T)+\rho(T,T')}{1-\rho(T,T')} \\ &\leq \frac{\frac{3\xi(T)}{2}}{1-\frac{\xi(T)}{2}} \\ &\leq 3\xi(T).\end{aligned}$$

Here we used the property that $\xi(T) \leq 1$ in obtaining the final inequality. Consequently, noting that $\gamma \in [\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}]$ implies that

$$\chi(\Omega,T',\gamma) = \max\left\{\frac{\xi(T')}{\gamma}, 2\mu(\Omega)\gamma\right\} \leq \frac{\nu\alpha}{\beta(2-\nu)}. \tag{C.1}$$

Part 1: The proof of this step proceeds in a similar manner to that of Lemma 3.5. First we have for $S \in \Omega, L \in T'$ with $\|S\|_\infty = \gamma, \|L\|_2 = 1$:

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{I}^*(S+L)\|_\infty &\geq \|\mathcal{P}_\Omega\mathcal{I}^*S\|_\infty - \|\mathcal{P}_\Omega\mathcal{I}^*L\|_\infty \\ &\geq \alpha\gamma - \|\mathcal{I}^*L\|_\infty \\ &\geq \alpha\gamma - \beta\xi(T').\end{aligned}$$

Next under the same conditions on S, L,

$$\begin{aligned}\|\mathcal{P}_{T'}\mathcal{I}^*(S+L)\|_2 &\geq \|\mathcal{P}_{T'}\mathcal{I}^*L\|_2 - \|\mathcal{P}_{T'}\mathcal{I}^*S\|_2 \\ &\geq \alpha - 2\|\mathcal{I}^*S\|_2 \\ &\geq \alpha - 2\beta\mu(\Omega)\gamma.\end{aligned}$$

Combining these last two bounds with (C.1), we conclude that

$$\begin{aligned}\min_{(S,L)\in\mathcal{Y},\ \|S\|_\infty=\gamma,\ \|L\|_2=1} g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(S,L)) &\geq \alpha-\beta\max\left\{\frac{\xi(T')}{\gamma},2\mu(\Omega)\gamma\right\}\\ &\geq \alpha-\frac{\nu\alpha}{2-\nu}\\ &= \frac{2\alpha(1-\nu)}{2-\nu}\\ &\geq \frac{\alpha}{2},\end{aligned}$$

where the final inequality follows from the assumption that $\nu\in(0,\frac{1}{2}]$.

Part 2: Note that for $S\in\Omega, L\in T'$ with $\|S\|_\infty\leq\gamma, \|L\|_2\leq 1$

$$\begin{aligned}\|\mathcal{P}_{\Omega^\perp}\mathcal{I}^*(S+L)\|_\infty &\leq \|\mathcal{P}_{\Omega^\perp}\mathcal{I}^*S\|_\infty+\|\mathcal{P}_{\Omega^\perp}\mathcal{I}^*L\|_\infty\\ &\leq \delta\gamma+\beta\xi(T').\end{aligned}$$

Similarly

$$\begin{aligned}\|\mathcal{P}_{T'^\perp}\mathcal{I}^*(S+L)\|_2 &\leq \|\mathcal{P}_{T'^\perp}\mathcal{I}^*S\|_2+\|\mathcal{P}_{T'^\perp}\mathcal{I}^*L\|_2\\ &\leq \beta\gamma\mu(\Omega)+\delta.\end{aligned}$$

Combining these last two bounds with the bounds from the first part, we have that

$$\begin{aligned}\left\|\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\left(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\right)^{-1}\right\|_{g_\gamma\to g_\gamma} &\leq \frac{\delta+\beta\max\left\{\frac{\xi(T')}{\gamma},2\mu(\Omega)\gamma\right\}}{\alpha-\beta\max\left\{\frac{\xi(T')}{\gamma},2\mu(\Omega)\gamma\right\}}\\ &\leq \frac{\delta+\frac{\nu\alpha}{2-\nu}}{\alpha-\frac{\nu\alpha}{2-\nu}}\\ &\leq \frac{(1-2\nu)\alpha+\frac{\nu\alpha}{2-\nu}}{\alpha-\frac{\nu\alpha}{2-\nu}}\\ &= 1-\nu.\end{aligned}$$

This concludes the proof of the proposition. □

D Proof of main result

Here we prove Theorem 4.1. Throughout this section we denote $m=\max\{1,\frac{1}{\gamma}\}$. Further $\Omega=\Omega(K_O^*)$ and $T=T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the tangent spaces at the "true" sparse matrix $S^*=K_O^*$ and low-rank matrix $L^*=K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$. We assume that

$$\gamma\in\left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha},\frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right] \tag{D.1}$$

We also let $E_n=\Sigma_O^n-\Sigma_O^*$ denote the difference between the true marginal covariance and the sample covariance. Finally we let $D=\max\{1,\frac{\nu\alpha}{3\beta(2-\nu)}\}$ throughout this section. For γ in the above range we note that

$$m\leq\frac{D}{\xi(T)}. \tag{D.2}$$

Standard facts that we use throughout this section are that $\xi(T) \leq 1$ and that $\|M\|_\infty \leq \|M\|_2$ for any matrix M.

We study the following convex program:

$$\begin{aligned}(\bar{S}_n, \bar{L}_n) = \arg\min_{S,L} \; & \operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0.\end{aligned} \tag{D.3}$$

Comparing (D.3) with the convex program (4.1), the main difference is that we do not constraint the variable L to be positive semidefinite in (D.3) (recall that the nuclear norm of a positive semidefinite matrix is equal to its trace). However we show that the unique optimum $(\bar{S}_n, \bar{L}_n)$ of (D.3) under the hypotheses of Theorem 4.1 is such that $\bar{L}_n \succeq 0$ (with high probability). Therefore we conclude that $(\bar{S}_n, \bar{L}_n)$ is also the unique optimum of (4.1). The subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = UDV^T$ is as follows:

$$N \in \partial\|M\|_* \quad \Leftrightarrow \quad \mathcal{P}_{T(M)}(N) = UV^T, \;\; \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \leq 1.$$

The proof of this theorem consists of a number of steps, each of which is analyzed in separate sections below. We explicitly keep track of the constants α, β, ν, ψ. The key ideas are as follows:

1. We show that if we solve the convex program (D.3) subject to the additional constraints that $S \in \Omega$ and $L \in T'$ for some T' "close to" T (measured by $\rho(T', T)$), then the error between the optimal solution $(\bar{S}_n, \bar{L}_n)$ and the underlying matrices (S^*, L^*) is small. This result is discussed in Appendix D.2.

2. We analyze the optimization problem (D.3) with the additional constraint that the variables S and L belong to the algebraic varieties of sparse and low-rank matrices respectively, and that the corresponding tangent spaces are close to the tangent spaces at (S^*, L^*). We show that under suitable conditions on the minimum nonzero singular value of the true low-rank matrix L^* and on the minimum magnitude nonzero entry of the true sparse matrix S^*, the optimum of this modified program is achieved at a *smooth* point of the underlying varieties. In particular the bound on the minimum nonzero singular value of L^* helps bound the curvature of the low-rank matrix variety locally around L^* (we use the results described in Appendix B). These results are described in Appendix D.3.

3. The next step is to show that the variety constraint can be linearized and changed to a tangent-space constraint (see Appendix D.4), thus giving us a *convex program*. Under suitable conditions this tangent-space constrained program also has an optimum that has the same support/rank as the true (S^*, L^*). Based on the previous step these tangent spaces in the constraints are close to the tangent spaces at the true (S^*, L^*). Therefore we use the first step to conclude that the resulting error in the estimate is small.

4. Finally we show that under the identifiability conditions of Section 3 these tangent-space constraints are inactive at the optimum (see Appendix D.7). Therefore we conclude with the statement that the optimum of the convex program (D.3) without any variety constraints is achieved at a pair of matrices that have the same support/rank as the true (S^*, L^*) (with high probability). Further the low-rank component of the solution is positive semidefinite, thus allowing us to conclude that the original convex program (4.1) also provides estimates that are consistent.

D.1 Bounded curvature of matrix inverse

Consider the Taylor series of the inverse of a matrix:

$$(M+\Delta)^{-1} = M^{-1} - M^{-1}\Delta M^{-1} + R_{M^{-1}}(\Delta),$$

where

$$R_{M^{-1}}(\Delta) = M^{-1}\left[\sum_{k=2}^{\infty}(-\Delta M^{-1})^k\right].$$

This infinite sum converges for Δ sufficiently small. The following proposition provides a bound on the second-order term specialized to our setting:

Proposition D.1. *Suppose that γ is in the range given by (D.1). Let $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$ for $C_1 = \psi(1+\frac{\alpha}{6\beta})$, and for any (Δ_S, Δ_L) with $\Delta_S \in \Omega$. Then we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))) \leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)}.$$

Proof: We have that

$$\begin{aligned}
\|\mathcal{A}(\Delta_S, \Delta_L)\|_2 &\leq \|\Delta_S\|_2 + \|\Delta_L\|_2 \\
&\leq \gamma\mu(\Omega)\frac{\|\Delta_S\|_\infty}{\gamma} + \|\Delta_L\|_2 \\
&\leq (1+\gamma\mu(\Omega))g_\gamma(\Delta_S, \Delta_L) \\
&\leq (1+\frac{\alpha}{6\beta})g_\gamma(\Delta_S, \Delta_L) \\
&\leq \frac{1}{2\psi},
\end{aligned}$$

where the second-to-last inequality follows from the range for γ (D.1) and that $\nu \in (0, \frac{1}{2}]$, and the final inequality follows from the bound on $g_\gamma(\Delta_S, \Delta_L)$. Therefore,

$$\begin{aligned}
\|R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 &\leq \psi\sum_{k=2}^{\infty}(\|\Delta_S + \Delta_L\|_2\psi)^k \\
&\leq \psi^3\|\Delta_S + \Delta_L\|_2^2\frac{1}{1-\|\Delta_S+\Delta_L\|_2\psi} \\
&\leq 2\psi^3(1+\frac{\alpha}{6\beta})^2 g_\gamma(\Delta_S, \Delta_L)^2 \\
&= 2\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2.
\end{aligned}$$

Here we apply the last two inequalities from above. Since the $\|\cdot\|_\infty$-norm is bounded above by the spectral norm $\|\cdot\|_2$, we have the desired result. □

D.2 Bounded errors

Next we analyze the following convex program subject to certain additional tangent-space constraints:

$$\begin{aligned}
(\hat{S}_\Omega, \hat{L}_{T'}) = \arg\min_{S,L}\ &\mathrm{tr}[(S-L)\ \Sigma_O^n] - \log\det(S-L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\
\text{s.t.}\quad &S - L \succ 0,\quad S \in \Omega,\quad L \in T',
\end{aligned} \tag{D.4}$$

for some subspace T'. We show that if T' is any tangent space to the low-rank matrix variety such that $\rho(T,T') \leq \frac{\xi(T)}{2}$, then we can bound the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})$. Let $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$ denote the normal component of the true low-rank matrix at T', and recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the true marginal covariance and the sample covariance. The proof of the following result uses Brouwer's fixed-point theorem [28], and is inspired by the proof of a similar result in [29] for standard sparse graphical model recovery without latent variables.

Proposition D.2. *Let the error (Δ_S, Δ_L) in the solution of the convex program (D.4) (with T' such that $\rho(T',T) \leq \frac{\xi(T)}{2}$) be as defined above. Further let $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, and define*

$$r = \max\left\{\frac{8}{\alpha}\left[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) + \lambda_n\right], \|\mathcal{C}_{T'}\|_2\right\}.$$

If we have that

$$r \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\},$$

for γ in the range given by (D.1), then

$$g_\gamma(\Delta_S, \Delta_L) \leq 2r.$$

Proof: Based on Proposition 3.6 we note that the convex program (D.4) is strictly convex (because the negative log-likelihood term has a strictly positive-definite Hessian due to the constraints involving transverse tangent spaces), and therefore the optimum is unique. Applying the optimality conditions of the convex program (D.4) at the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T'^\perp} \in T'^\perp$ such that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{\Omega^\perp} \in -\lambda_n\gamma\partial\|\hat{S}_\Omega\|_1, \quad \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{T'^\perp} \in \lambda_n\partial\|\hat{L}_{T'}\|_*.$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T'$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_\Omega, \quad \mathcal{P}_{T'}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_{T'},$$

where $Z_\Omega \in \Omega, Z_{T'} \in T'$ and $\|Z_\Omega\|_\infty = \lambda_n\gamma, \|Z_{T'}\|_2 \leq 2\lambda_n$ (we use here the fact that projecting onto a tangent space T' increases the spectral norm by at most a factor of two). Denoting $Z = [Z_\Omega, Z_{T'}]$, we conclude that

$$\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z, \tag{D.5}$$

with $g_\gamma(Z) \leq 2\lambda_n$. Since the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$ is unique, one can check using Lagrangian duality theory [31] that $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique solution of the equation (D.5). Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}$ in terms of the errors (Δ_S, Δ_L), we have using the Taylor series of the matrix inverse that

$$\begin{aligned}
\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} &= \Sigma_O^n - [\mathcal{A}(\Delta_S, \Delta_L) + (\Sigma_O^*)^{-1}]^{-1} \\
&= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}(\Delta_S, \Delta_L) \\
&= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L) + \mathcal{I}^*\mathcal{C}_{T'}.
\end{aligned} \tag{D.6}$$

Since T' is a tangent space such that $\rho(T',T) \leq \frac{\xi(T)}{2}$, we have from Proposition 3.6 that the operator $\mathcal{B} = \left(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}\right)^{-1}$ from $\mathcal{Y}$ to $\mathcal{Y}$ is bijective and is well-defined. Now consider the following matrix-valued function from $(\delta_S, \delta_L) \in \mathcal{Y}$ to $\mathcal{Y}$:

$$F(\delta_S, \delta_L) = (\delta_S, \delta_L) - \mathcal{B}\left\{\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\delta_S, \delta_L) + \mathcal{I}^*\mathcal{C}_{T'}] - Z\right\}.$$

A point $(\delta_S, \delta_L) \in \mathcal{Y}$ is a fixed-point of F if and only if $\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^*\mathcal{C}_{T'}] = Z$. Applying equations (D.5) and (D.6) above, we then see that the only fixed-point of F by construction is the "true" error $\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)$ restricted to $\mathcal{Y}$. The reason for this is that, as discussed above, $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique optimum of (D.4) and therefore is the *unique solution* of (D.5). Next we show that this unique fixed-point of F lies in the ball $\mathbb{B}_r = \{(\delta_S, \delta_L) \mid g_\gamma(\delta_S, \delta_L) \leq r, \quad (\delta_S, \delta_L) \in \mathcal{Y}\}$.

In order to prove this step, we resort to Brouwer's fixed point theorem [28]. In particular we show that the function F maps the ball $\mathbb{B}_r$ onto itself. Since F is a continuous function and $\mathbb{B}_r$ is a compact set, we can conclude the proof of this proposition. Simplifying the function F, we have that

$$F(\delta_S, \delta_L) = \mathcal{B}\left\{\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) - \mathcal{I}^*\mathcal{C}_{T'}] + Z\right\}.$$

Consequently, we have from Proposition 3.6 that

$$\begin{aligned} g_\gamma(F(\delta_S, \delta_L)) &\leq \frac{2}{\alpha}\, g_\gamma\left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{C}_{T'}] - Z\right) \\ &\leq \frac{4}{\alpha}\, \left\{g_\gamma(\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{C}_{T'}]) + \lambda_n\right\} \\ &\leq \frac{r}{2} \;+\; \frac{4}{\alpha}\, g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))), \end{aligned}$$

where in the second inequality we use the fact that $g_\gamma(\mathcal{P}_{\mathcal{Y}}(\cdot, \cdot)) \leq 2g_\gamma(\cdot, \cdot)$ and that $g_\gamma(Z) \leq 2\lambda_n$, and in the final inequality we use the assumption on r.

We now bound the term $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L)))$ using Proposition D.1 as $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$:

$$\begin{aligned} \frac{4}{\alpha}\, g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))) &\leq \frac{8D\psi C_1^2(g_\gamma(\delta_S, \delta_L) + \|\mathcal{C}_{T'}\|_2)^2}{\xi(T)\alpha} \\ &\leq \frac{32D\psi C_1^2 r^2}{\xi(T)\alpha} \\ &\leq \frac{32D\psi C_1^2 r}{\xi(T)\alpha}\, \frac{\alpha\xi(T)}{64D\psi C_1^2} \\ &\leq \frac{r}{2}, \end{aligned}$$

where we have used the fact that $r \leq \frac{\alpha\xi(T)}{64D\psi C_1^2}$. Hence $g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq r$ by Brouwer's fixed-point theorem. Finally we observe that

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) + \|\mathcal{C}_{T'}\|_2 \\ &\leq 2r. \end{aligned}$$

□

D.3 Solving a variety-constrained problem

In order to prove that the solution $(\bar{S}_n, \bar{L}_n)$ of (D.3) has the same sparsity pattern/rank as (S^*, L^*), we will study an optimization problem that explicitly enforces these constraints. Specifically, we consider the following *non-convex* constraint set:

$$\begin{aligned} \mathcal{M} \;=\; &\{(S, L) \mid S \in \Omega(S^*),\ \mathrm{rank}(L) \leq \mathrm{rank}(L^*), \\ &\|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2},\ g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n\} \end{aligned}$$

Recall that $S^* = K^*_O$ and $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$. The first constraint ensures that the tangent space at S is the same as the tangent space at S^*; therefore the support of S is contained in the support of S^*. The second and third constraints ensure that L lives in the appropriate low-rank variety, but has a tangent space "close" to the tangent space T. The final constraint roughly bounds the sum of the errors $(S - S^*) + (L^* - L)$; note that this does not necessarily bound the individual errors. Notice that the only non-convex constraint is that $\text{rank}(L) \leq \text{rank}(L^*)$. We then have the following nonlinear program:

$$\begin{aligned}(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}}) = \arg\min_{S,L} \ & \text{tr}[(S - L)\ \Sigma^n_O] - \log\det(S - L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0, \quad (S, L) \in \mathcal{M}.\end{aligned} \tag{D.7}$$

Under suitable conditions this nonlinear program is shown to have a unique solution. Each of the constraints in $\mathcal{M}$ is useful for proving the consistency of the solution of the convex program (D.3). We show that under suitable conditions the constraints in $\mathcal{M}$ are actually inactive at the optimal $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$, thus allowing us to conclude that the solution of (D.3) is also equal to $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$; hence the solution of (D.3) shares the consistency properties of $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$. A number of interesting properties can be derived simply by studying the constraint set $\mathcal{M}$.

Proposition D.3. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *For* γ *in the range specified by* (D.1) *and letting* $C_2 = \frac{48}{\alpha} + \frac{1}{\psi^2}$, *we have that* $g_\gamma(\Delta_S, \Delta_L) \leq C_2\lambda_n$.

Proof: We have by the triangle inequality that

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\mathcal{P}_\Omega(\Delta_S), \mathcal{P}_T(\Delta_L))) &\leq 11\lambda_n + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\mathcal{P}_{\Omega^\perp}(\Delta_S), \mathcal{P}_{T^\perp}(\Delta_L))) \\ &\leq 11\lambda_n + m\psi^2\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq 12\lambda_n, \end{aligned}$$

as $m \leq \frac{D}{\xi(T)}$. Therefore, we have that $g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq 24\lambda_n$, where $\mathcal{Y} = \Omega \times T$. Consequently, we can apply Proposition 3.6 to conclude that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq \frac{48\lambda_n}{\alpha}.$$

Finally, we use the triangle inequality again to conclude that

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}(\Delta_S, \Delta_L)) \\ &\leq \frac{48\lambda_n}{\alpha} + m\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq C_2\lambda_n. \end{aligned}$$

□

This simple result immediately leads to a number of useful corollaries. For example we have that under a suitable bound on the minimum nonzero singular value of $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, the constraint in $\mathcal{M}$ along the normal direction $T^\perp$ is locally inactive. Next we list several useful consequences of Proposition D.3.

Corollary D.4. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *Suppose* γ *is in the range specified by* (D.1), *and let* $C_3 = \left(\frac{6(2-\nu)}{\nu} + 1\right) C_2^2\psi^2 D$ *and* $C_4 = C_2 + \frac{3\alpha C_2^2(2-\nu)}{16(3-\nu)}$ *(where* C_2 *is as defined in Proposition D.3). Let the minimum nonzero singular value* σ *of* $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ *be such that* $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ *for* $C_5 = \max\{C_3, C_4\}$, *and suppose that the smallest magnitude nonzero entry of* S^* *is greater than* $\frac{C_6\lambda_n}{\mu(\Omega)}$ *for* $C_6 = \frac{C_2\nu\alpha}{\beta(2-\nu)}$. *Setting* $T' = T(L)$ *and* $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$, *we then have that:*

1. *L has rank equal to* $\operatorname{rank}(L^*)$*, i.e., L is a smooth point of the variety of matrices with rank less than or equal to* $\operatorname{rank}(L^*)$*. In particular L has the same inertia as L^*.*

2. $\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \leq \frac{\xi(T)\lambda_n}{19D\psi^2}$.

3. $\rho(T, T') \leq \frac{\xi(T)}{4}$.

4. $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$.

5. $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.

6. $\operatorname{sign}(S) = \operatorname{sign}(S^*)$.

Proof: We note the following facts before proving each step. First $C_2 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2} \geq \frac{\xi(T)}{D\psi^2}$. Second $\xi(T) \leq 1$. Third we have from Proposition D.3 that $\|\Delta_L\|_2 \leq C_2\lambda_n$. Finally $\frac{6(2-\nu)}{\nu} \geq 18$ for $\nu \in (0, \frac{1}{2}]$. We prove each step separately.

For the first step, we note that

$$\sigma \geq \frac{C_3\lambda_n}{\xi(T)^2} \geq \frac{19C_2^2\psi^2 D\lambda_n}{\xi(T)^2} \geq \frac{19C_2\lambda_n}{\xi(T)} \geq 8C_2\lambda_n \geq 8\|\Delta_L\|_2.$$

Hence L is a smooth point with rank equal to $\operatorname{rank}(L^*)$, and specifically has the same inertia as L^*.

For the second step, we use the fact that $\sigma \geq 8\|\Delta_L\|_2$ to apply Proposition 2.2:

$$\|\mathcal{P}_{T^\perp}(\Delta_L)\| \leq \frac{\|\Delta_L\|_2^2}{\sigma} \leq \frac{C_2^2\xi(T)^2\lambda_n^2}{C_3\lambda_n} \leq \frac{\xi(T)\lambda_n}{19D\psi^2}.$$

For the third step we apply Proposition 2.1 (by using the conclusion from above that $\sigma \geq 8\|\Delta_L\|_2$) so that

$$\rho(T, T') \leq \frac{2\|\Delta_L\|_2}{\sigma} \leq \frac{2C_2\xi(T)^2}{C_3} \leq \frac{2\xi(T)^2}{19C_2D\psi^2} \leq \frac{\xi(T)}{4}.$$

For the fourth step let σ' denote the minimum singular value of L. Consequently,

$$\sigma' \geq \frac{C_3\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq C_2\lambda_n\left[\frac{19C_2D\psi^2}{\xi(T)^2} - 1\right] \geq 8\|\Delta_L\|_2.$$

Using the same reasoning as in the proof of the second step, we have that

$$\begin{aligned}\|\mathcal{C}_{T'}\|_2 &\leq \frac{\|\Delta_L\|_2^2}{\sigma'} \leq \frac{C_2^2\lambda_n^2}{(\frac{C_3}{\xi(T)^2} - C_2)\lambda_n} = \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2D\psi^2(\frac{6(2-\nu)}{\nu}) + C_2^2D\psi^2 - C_2\xi(T)^2} \\ &\leq \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2D\psi^2(\frac{6(2-\nu)}{\nu})} \leq \frac{\nu\xi(T)\lambda_n}{6(2-\nu)D\psi^2}.\end{aligned}$$

Hence

$$g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) \leq m\psi^2\|\mathcal{C}_{T'}\|_2 \leq \frac{\lambda_n\nu}{6(2-\nu)}.$$

For the fifth step the bound on σ' implies that

$$\sigma' \geq \frac{C_4\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq \frac{3C_2^2\alpha(2-\nu)}{16(3-\nu)}\lambda_n$$

Since $\sigma' \geq 8\|\Delta_L\|_2$, we have from Proposition 2.2 and some algebra that

$$\|\mathcal{C}_{T'}\|_2 \leq \frac{C_2^2\lambda_n^2}{\sigma'} \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}.$$

For the final step since $\|\Delta_S\|_\infty \leq \gamma C_2 \lambda_n$, the assumed lower bound on the minimum magnitude nonzero entry of S^* guarantees that $\text{sign}(S) = \text{sign}(S^*)$. $\square$

Notice that this corollary applies to *any* $(S, L) \in \mathcal{M}$, and is hence applicable to *any solution* $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ of the $\mathcal{M}$-constrained program (D.7). For now we choose an arbitrary solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ and proceed. In the next steps we show that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is *the unique* solution to the convex program (D.3), thus showing that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is also the unique solution to (D.7).

D.4 From variety constraint to tangent-space constraint

Given the solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$, we show that the solution to the convex program (D.4) with the tangent space constraint $L \in T_\mathcal{M} \triangleq T(\hat{L}_\mathcal{M})$ is the same as $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ under suitable conditions:

$$\begin{aligned}(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}}) = \arg\min_{S,L} \ & \text{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \ + \ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0, \quad S \in \Omega, \quad L \in T_\mathcal{M}.\end{aligned} \tag{D.8}$$

Assuming the bound of Corollary D.4 on the minimum singular value of L^* the uniqueness of the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ is assured. This is because we have from Proposition 3.6 and from Corollary D.4 that $\mathcal{I}^*$ is injective on $\Omega \oplus T_\mathcal{M}$. Therefore the Hessian of the convex objective function of (D.8) is strictly positive-definite at $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$.

We let $\mathcal{C}_\mathcal{M} = \mathcal{P}_{T_\mathcal{M}^\perp}(L^*)$. Recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the sample covariance matrix and the marginal covariance matrix of the observed variables.

Proposition D.5. *Let γ be in the range specified by (D.1). Suppose that the minimum nonzero singular value σ of $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ is such that $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ (C_5 is defined in Corollary D.4). Suppose also that the minimum magnitude nonzero entry of S^* is greater than or equal to $\frac{C_6\lambda_n}{\mu(\Omega)}$ (C_6 is defined in Corollary D.4). Let $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Further suppose that*

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.$$

Then we have that

$$(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}}) = (\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M}).$$

Proof: Note first that the condition on the minimum singular value of L^* in Corollary D.4 is satisfied. Therefore we proceed with the following two steps:

1. First we can change the non-convex constraint $\text{rank}(L) \leq \text{rank}(L^*)$ to the linear constraint $L \in T(\hat{L}_\mathcal{M})$. This is because the lower bound assumed for σ implies that $\hat{L}_\mathcal{M}$ is a smooth point of the algebraic variety of matrices with rank less than or equal to $\text{rank}(L^*)$ (from Corollary D.4). Due to the convexity of all the other constraints and the objective, the optimum of this "linearized" convex program will still be $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$.

2. Next we can again apply Corollary D.4 (based on the bound on σ) to conclude that the constraint $\|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2}$ is *locally inactive* at the point $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$.

Consequently, we have that $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$ can be written as the solution of a *convex program*:

$$\begin{aligned}(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}}) = \arg\min_{S,L} \ & \mathrm{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \ + \ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0, \quad S \in \Omega, \quad L \in T_{\mathcal{M}}, \\ & g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n.\end{aligned} \tag{D.9}$$

We now need to argue that the constraint $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n$ is also inactive in the convex program (D.9). We proceed by showing that the solution $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ of the convex program (D.8) has the property that $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_{\mathcal{M}}})) < 11\lambda_n$, which concludes the proof of this proposition. We have from Corollary D.4 that $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$. Since $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n \nu}{6(2-\nu)}$ by assumption, one can verify that

$$\begin{aligned}\frac{8}{\alpha}\left[\lambda_n + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}})\right] &\leq \frac{8\lambda_n}{\alpha}\left[1 + \frac{\nu}{3(2-\nu)}\right] \\ &= \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)} \\ &\leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64 D\psi C_1^2}\right\}.\end{aligned}$$

The last line follows from the assumption on λ_n. We also note that $\|C_{T_{\mathcal{M}}}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Corollary D.4, which implies that $\|C_{T_{\mathcal{M}}}\|_2 \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64 D\psi C_1^2}\right\}$. Letting $(\Delta_S, \Delta_L) = (S_\Omega - S^*, L^* - \hat{L}_{T_{\mathcal{M}}})$, we can conclude from Proposition D.2 that $g_\gamma(\Delta_L, \Delta_S) \leq \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}$. Next we apply Proposition D.1 (as $g_\gamma(\Delta_L, \Delta_S) \leq \frac{1}{2C_1}$) to conclude that

$$\begin{aligned}g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) &\leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)} \\ &\leq \frac{2D\psi C_1^2}{\xi(T)} \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)} \frac{\alpha\xi(T)}{32 D\psi C_1^2} \\ &\leq \frac{2(3-\nu)\lambda_n}{3(2-\nu)}.\end{aligned} \tag{D.10}$$

From the optimality conditions of (D.8) one can also check that for $\mathcal{Y} = \Omega \times T_{\mathcal{M}}$,

$$\begin{aligned}g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) &\leq 2\lambda_n + g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}) + g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger E_n) \\ &\leq 2[\lambda_n + g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) \\ &\quad + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}})] \\ &\leq 4\left[\frac{2(3-\nu)\lambda_n}{3(2-\nu)}\right].\end{aligned}$$

Here we used (D.10) in the last inequality, and also that $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$ (as noted above from Corollary D.4) and that $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n \nu}{6(2-\nu)}$. Therefore,

$$g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq \frac{16\lambda_n}{3}, \tag{D.11}$$

because $\nu \in (0, \frac{1}{2}]$. Based on Proposition 3.6 (the second part), we also have that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq (1-\nu)\frac{16\lambda_n}{3} \leq \frac{16\lambda_n}{3}. \tag{D.12}$$

Summarizing steps (D.11) and (D.12),

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\Delta_S, \Delta_L)) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}}) \\ &\leq \frac{16\lambda_n}{3} + \frac{16\lambda_n}{3} + \frac{\lambda_n\nu}{6(2-\nu)} \\ &\leq \frac{32\lambda_n}{3} + \frac{\lambda_n}{18} \\ &< 11\lambda_n. \end{aligned}$$

This concludes the proof of the proposition. □

This proposition has the following important consequence.

Corollary D.6. *Under the assumptions of Proposition D.5 we have that* $\mathrm{rank}(\hat{L}_{T_{\mathcal{M}}}) = \mathrm{rank}(L^*)$ *and that* $T(\hat{L}_{T_{\mathcal{M}}}) = T_{\mathcal{M}}$. *Moreover,* $\hat{L}_{T_{\mathcal{M}}}$ *actually has the same inertia as* L^*. *We also have that* $\mathrm{sign}(\hat{S}_\Omega) = \mathrm{sign}(S^*)$.

D.5 Removing the tangent-space constraints

The following lemma provides a simple set of sufficient conditions under which the optimal solution $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ of (D.8) satisfies the optimality conditions of the convex program (D.3) (without the tangent space constraints).

Lemma D.7. *Let* $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ *be the solution to the tangent-space constrained convex program* (D.8). *Suppose that the assumptions of Proposition D.5 hold. If in addition we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))) \leq \frac{\lambda_n\nu}{6(2-\nu)},$$

then $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ *is also the unique optimum of the convex program* (D.3).

Proof: Recall from Corollary D.6 that the tangent space at $\hat{L}_{T_{\mathcal{M}}}$ is equal to $T_{\mathcal{M}}$. Applying the optimality conditions of the convex program (D.8) at the optimum $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T_{\mathcal{M}}^\perp} \in T_{\mathcal{M}}^\perp$ such that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} + Q_{\Omega^\perp} \in -\lambda_n\gamma\partial\|\hat{S}_\Omega\|_1, \quad \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} + Q_{T_{\mathcal{M}}^\perp} \in \lambda_n\partial\|\hat{L}_{T_{\mathcal{M}}}\|_*.$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T_{\mathcal{M}}$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = -\lambda_n\gamma\mathrm{sign}(S^*), \quad \mathcal{P}_{T_{\mathcal{M}}}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = \lambda_n UV^T,$$

where $\hat{L}_{T_{\mathcal{M}}} = UDV^T$ is a reduced SVD of $\hat{L}_{T_{\mathcal{M}}}$. Denoting $Z = [-\lambda_n\gamma\mathrm{sign}(S^*), \lambda_n UV^T]$, we conclude that

$$\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = Z, \tag{D.13}$$

with $g_\gamma(Z) = \lambda_n$. It is clear that the optimality condition of the convex program (D.3) (without the tangent-space constraints) on $\mathcal{Y}$ is satisfied. All we need to show is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}]) < \lambda_n. \tag{D.14}$$

Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_\mathcal{M}})^{-1}$ in terms of the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})$, we have that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_\mathcal{M}})^{-1} = E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}(\Delta_S, \Delta_L).$$

Restating the condition (D.13) on $\mathcal{Y}$, we have that

$$\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L) = Z + \mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}]. \tag{D.15}$$

(Recall that $\mathcal{C}_{T_\mathcal{M}} = \mathcal{P}_{T_\mathcal{M}^\perp}(L^*)$.) A sufficient condition to show (D.14) and complete the proof of this lemma is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) < \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}]).$$

We prove this inequality next. Recall from Corollary D.4 that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Therefore, from equation (D.15) we can conclude that

$$\begin{aligned} g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) &\leq \lambda_n + 2(g_\gamma(\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}])) \\ &\leq \lambda_n + 2\left[\frac{3\lambda_n\nu}{6(2-\nu)}\right] \\ &= \frac{2\lambda_n}{2-\nu}. \end{aligned}$$

Here we used the bounds assumed on $g_\gamma(\mathcal{A}^\dagger E_n)$ and on $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)))$.

Applying the second part of Proposition 3.6, we have that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) &\leq \frac{2\lambda_n(1-\nu)}{2-\nu} \\ &= \lambda_n - \frac{\nu\lambda_n}{2-\nu} \\ &< \lambda_n - \frac{\nu\lambda_n}{2(2-\nu)} \\ &\leq \lambda_n - g_\gamma(\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}]) \\ &\leq \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}]). \end{aligned}$$

Here the second-to-last inequality follows from the bounds on $g_\gamma(\mathcal{A}^\dagger E_n)$, $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)))$, and $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})$, and the last inequality follows from Lemma 3.4. This concludes the proof of the lemma. □

D.6 Probabilistic analysis

All the analysis described so far in this section has been completely deterministic in nature. Here we present the probabilistic component of our proof. Specifically, we study the rate at which the sample covariance matrix converges to the true covariance matrix. The following result from [10] plays a key role in our analysis:

Theorem D.8. *Given natural numbers n, p with $p \leq n$, let Γ be a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. Then the largest and smallest singular values $s_1(\Gamma)$ and $s_p(\Gamma)$ of Γ are such that*

$$\max\left\{\Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + t\right], \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - t\right]\right\} \leq \exp\left\{-\tfrac{nt^2}{2}\right\},$$

for any $t > 0$.

Using this result the next lemma provides a probabilistic bound between the sample covariance Σ_O^n formed using n samples and the true covariance Σ_O^* in spectral norm. This result is well-known, and we mainly discuss it here for completeness and also to show explicitly the dependence on $\psi = \|\Sigma_O^*\|_2$ defined in (3.6).

Lemma D.9. *Let $\psi = \|\Sigma_O^*\|_2$. Given any $\delta > 0$ with $\delta \leq 8\psi$, let the number of samples n be such that $n \geq \frac{64p\psi^2}{\delta^2}$. Then we have that*

$$\Pr\left[\|\Sigma_O^n \ - \ \Sigma_O^*\|_2 \geq \delta\right] \leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.$$

Proof: Since the spectral norm is unitarily invariant, we can assume that Σ_O^* is diagonal without loss of generality. Let $\bar{\Sigma}^n = (\Sigma_O^*)^{-\frac{1}{2}}\Sigma_O^n(\Sigma_O^*)^{-\frac{1}{2}}$, and let $s_1(\bar{\Sigma}^n), s_p(\bar{\Sigma}^n)$ denote the largest/smallest singular values of $\bar{\Sigma}^n$. Note that $\bar{\Sigma}^n$ can be viewed as the sample covariance matrix formed from n independent samples drawn from a model with identity covariance, i.e., $\bar{\Sigma}^n = \Gamma\Gamma^T$ where Γ denotes a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. We then have that

$$\begin{aligned}
\Pr\left[\|\Sigma_O^n \ - \ \Sigma_O^*\|_2 \geq \delta\right] &\leq \Pr\left[\|\bar{\Sigma}^n \ - I\|_2 \geq \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\bar{\Sigma}^n) \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\bar{\Sigma}^n) \leq 1 - \tfrac{\delta}{\psi}\right] \\
&= \Pr\left[s_1(\Gamma)^2 \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\Gamma)^2 \leq 1 - \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \tfrac{\delta}{4\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \tfrac{\delta}{4\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + \tfrac{\delta}{8\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - \tfrac{\delta}{8\psi}\right] \\
&\leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.
\end{aligned}$$

Here we used the fact that $n \geq \frac{64p\psi^2}{\delta^2}$ in the fourth inequality, and we applied Theorem D.8 to obtain the final inequality by setting $t = \frac{\delta}{8\psi}$. □

The following corollary describes relates the number of samples required for an error bound to hold with probability $1 - 2\exp\{-p\}$.

Corollary D.10. *Let Σ_O^n be the sample covariance formed from n samples of the observed variables. Set $\delta_n = \sqrt{\frac{128p\psi^2}{n}}$. If $n \geq 2p$, then we have with probability greater than $1 - 2\exp\{-p\}$ that*

$$\Pr\left[\|\Sigma_O^n \ - \ \Sigma_O^*\|_2 \leq \delta_n\right] \geq 1 - 2\exp\{-p\}.$$

Proof: We note that $n \geq 2p$ implies that $\delta_n \leq 8\psi$, and apply Lemma D.9. □

D.7 Putting it all together

In this section we tie together the results obtained thus far to conclude the proof of Theorem 4.1. We only need to show that the sufficient conditions of Lemma D.7 are satisfied. It follows directly from Corollary D.6 that the low-rank part $\hat{L}_{T_\mathcal{M}}$ is positive semidefinite, which implies that $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ is also the solution to the original regularized maximum-likelihood convex program (4.1) with the positive-semidefinite constraint. As usual set $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})$, and set $E_n = \Sigma_O^n - \Sigma_O^*$.

Assumptions: We specify here the constants that were suppressed in the statement of Theorem 4.1:

1. Let $C_7 = \frac{\alpha\nu}{32(3-\nu)D} \min\left\{\frac{1}{4C_1}, \frac{\alpha\nu}{256D(3-\nu)\psi C_1^2}\right\}$, and let the number of samples n be such that

$$n \geq \frac{p}{\xi(T)^4} \max\left\{\frac{128\psi^2}{C_7^2}, 2\right\}.$$

   Note that $n \gtrsim \frac{p}{\xi(T)^4}$.

2. Set $\delta_n = \sqrt{\frac{128p\psi^2}{n}}$, and then set λ_n as follows:

$$\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}.$$

   Note that $\lambda_n \asymp \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$.

3. Let the minimum nonzero singular value σ of L^* be such that

$$\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2},$$

   where C_5 is defined in Corollary D.4. Note that $\sigma \gtrsim \frac{1}{\xi(T)^3}\sqrt{\frac{p}{n}}$.

4. Let the minimum magnitude nonzero entry θ of S^* be such that

$$\theta \geq \frac{C_6\lambda_n}{\mu(\Omega)},$$

   where C_6 is defined in Corollary D.4. Note that $\theta \gtrsim \frac{1}{\xi(T)\mu(\Omega)}\sqrt{\frac{p}{n}}$.

Proof of Theorem 4.1: We condition on the event that $\|E_n\|_2 \leq \delta_n$, which holds with probability greater than $1 - 2\exp\{-p\}$ from Corollary D.10 as $n \geq 2p$ by assumption. We note that based on the bound on n, we also have that

$$\delta_n \leq \xi(T)^2\left[\frac{\alpha\nu}{32(3-\nu)D}\min\left\{\frac{1}{4C_1}, \frac{\alpha\nu}{256D(3-\nu)\psi C_1^2}\right\}\right].$$

In particular, these bounds imply that

$$\delta_n \leq \frac{\alpha\xi(T)\nu}{32(3-\nu)D}\min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\} \tag{D.16}$$

and that

$$\delta_n \leq \frac{\alpha^2\xi(T)^2\nu^2}{8192\psi C_1^2(3-\nu)^2D^2}. \tag{D.17}$$

Both these weaker bounds are used later.

Based on the assumptions above, the requirements of Lemma D.7 on the minimum nonzero singular value of L^* and the minimum magnitude nonzero entry of S^* are satisfied. We only need to verify the bounds on λ_n and $g_\gamma(\mathcal{A}^\dagger E_n)$ from Proposition D.5, and the bound on $g_\gamma(\mathcal{A}^\dagger R\mathcal{A}(\Delta_S, \Delta_L))$ from Lemma D.7.

First we verify the bound on λ_n. Based on the setting of λ_n above and the bound on δ_n from (D.16), we have that

$$\begin{aligned}\lambda_n &= \frac{6D(2-\nu)\delta_n}{\xi(T)\nu} \\ &\leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned}$$

Next we combine the facts that $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$, and that $\|E_n\|_2 \leq \delta_n$ to conclude that

$$g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{D\delta_n}{\xi(T)} = \frac{\lambda_n \nu}{6(2-\nu)}.$$

Finally we provide a bound on the remainder by applying Propositions D.2 and D.1, which would satisfy the last remaining condition of Lemma D.7. In order to apply Proposition D.2, we note that

$$\begin{aligned}\frac{8}{\alpha}\left[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_\mathcal{M}}) + \lambda_n\right] &\leq \frac{8}{\alpha}\left[\frac{\nu}{3(2-\nu)} + 1\right]\lambda_n \\ &= \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)} \\ &= \frac{32(3-\nu)D}{\alpha\xi(T)\nu}\delta_n \\ &\leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned} \tag{D.18}$$

In the first inequality we used the fact that $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ (from above) and that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})$ is similarly bounded (from Corollary D.4 due to the bound on σ). In the second equality we used the relation $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. In the final inequality we used the bound on δ_n from (D.16). This satisfies one of the requirements of Proposition D.2. The other condition on $\|\mathcal{C}_{T_\mathcal{M}}\|_2$ is also similarly satisfied due to the bound on σ from Corollary D.4. Specifically, we have that $\|\mathcal{C}_{T_\mathcal{M}}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Corollary D.4, and use the same sequence of inequalities as above to satisfy the second requirement of Proposition D.2. Thus we conclude from Proposition D.2 and from (D.18) that

$$g_\gamma(\Delta_S, \Delta_L) \leq \frac{64(3-\nu)D}{\alpha\xi(T)\nu}\delta_n. \tag{D.19}$$

This bound implies that $g_\gamma(\Delta_S, \Delta_L) \lesssim \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$, which proves the parametric consistency part of the theorem.

Since the bound (D.19) also satisfies the condition of Proposition D.1 (from the inequality

following (D.18) above we see that $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$), we have that

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger R(\Delta_S + \Delta_L)) &\leq \frac{2D\psi C_1^2}{\xi(T)} g_\gamma(\Delta_S, \Delta_L)^2 \\ &\leq \frac{2D\psi C_1^2}{\xi(T)} \left(\frac{64(3-\nu)D}{\alpha\xi(T)\nu}\right)^2 \delta_n^2 \\ &= \left[\frac{8192\psi C_1^2(3-\nu)^2 D^2}{\alpha^2\xi(T)^2\nu^2}\delta_n\right] \frac{D\delta_n}{\xi(T)} \\ &\leq \frac{D\delta_n}{\xi(T)} \\ &= \frac{\lambda_n \nu}{6(2-\nu)}. \end{aligned}$$

In the final inequality we used the bound (D.17) on δ_n, and in the final equality we used the relation $\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$. This concludes the algebraic consistency part of the theorem. □

References

[1] Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statistics.* **37** 3099–3132.

[2] Bach, F. (2008). Consistency of trace norm minimization. *J. Mach. Lear. Res..* **9** 1019–1048.

[3] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statistics.* **36** 199–227.

[4] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statistics.* **36** 2577–2604.

[5] Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization.* Cambridge University Press.

[6] Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory.* **52** 489–509.

[7] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. of Comput. Math..* **9** 717–772.

[8] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2009). Robust principal component analysis? *Preprint.*

[9] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2009). Rank-sparsity incoherence for matrix decomposition. *Preprint.*

[10] Davidson, K. R. and Szarek, S.J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces.* **I** 317–366.

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B.* **39** 1–38.

[12] Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Math.*. **59** 797–829.

[13] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Info. Theory*. **52** 1289–1306.

[14] Elidan, G., Nachman, I., and Friedman, N. (2007). "Ideal Parent" structure learning for continuous variable Bayesian networks. *J. Mach. Lear. Res.*. **8** 1799–1833.

[15] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statistics*. **36** 2717–2756.

[16] Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*. **147** 186–197.

[17] Fazel, M. (2002). Matrix rank minimization with applications. *PhD thesis, Dep. Elec. Eng., Stanford University*. 2002.

[18] Golub, G. H. and Van Loan, C. H. (1990). *Matrix computations*. The Johns Hopkins Univ. Press.

[19] Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis*. Cambridge University Press.

[20] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statistics*. **29** 295–327.

[21] Kato, T. (1995). *Perturbation theory for linear operators*. Springer.

[22] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statistics*. **37** 4254–4278.

[23] Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

[24] Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Analysis*. **88** 365–411.

[25] Löfberg, J. (2004). YALMIP: A Toolbox for Modeling and Optimization in MATLAB. *Proceedings of the CACSD Conference, Taiwan*. Available from http://control.ee.ethz.ch/ joloef/yalmip.php.

[26] Marcenko, V. A. and Pastur, L. A. (1967). Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb*. **1** 507–536.

[27] Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statistics*. **34** 1436–1462.

[28] Ortega, J. M. and Rheinboldt, W. G. (1970). *Iterative solution of nonlinear equations in several variables*. Academic Press.

[29] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Preprint*.

[30] Recht, B., Fazel, M., and Parrilo, P. A. (2009). Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review, to appear*.

[31] Rockafellar, R. T. (1996). *Convex Analysis*. Princeton University Press.

[32] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Elec. J. Statistics*. **2** 494–515.

[33] Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statistics*. **14** 138–150.

[34] K. C. Toh, M. J. Todd, and R. H. Tutuncu. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming*. Available from http://www.math.nus.edu.sg/ mattohkc/sdpt3.html.

[35] Varga, R. S. (2000). *Matrix iterative analysis*. Springer-Verlag.

[36] Wang, C., Sun, D. and Toh, K. C. (2009). Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *Preprint*.

[37] Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*. **170** 1039–1053.

[38] Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostat*. **10** 515–534.

[39] Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*. **90** 831–844.

[40] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Lear. Res.*. **7** 2541–2567.